ISSN 2713-3192
DOI 10.15622/ia.2024.23.6
http://ia.spcras.ru

ТОМ 23 № 6

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

INFORMATICS AND AUTOMATION

СПб ФИЦ РАН

Санкт-Петербург
2024

INFORMATICS AND AUTOMATION

Volume 23 № 6, 2024

Scientific and educational journal primarily specialized in computer science, automation, robotics, applied mathematics, interdisciplinary research
Founded in 2002

Founder and Publisher
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

Editor: A.S. Viktorova **Interpreter:** Ya.N. Berezina
Art editor: N.A. Dormidontova

Editorial office address
SPC RAS, 39 litera A , 14-th line V.O., St. Petersburg, 199178, Russia
e-mail: ia@spcras.ru, web: http://ia.spcras.ru

The journal is indexed in Scopus

The journal is published under the scientific-methodological supervision of Department for Nanotechnologies and Information Technologies of the Russian Academy of Sciences

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

Том 23 № 6, 2024

Научный, научно-образовательный журнал с базовой специализацией
в области информатики, автоматизации, робототехники, прикладной математики
и междисциплинарных исследований.
Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»
(СПб ФИЦ РАН)

Выпускающий редактор: А.С. Викторова **Переводчик:** Я.Н. Березина
Художественный редактор: Н.А. Дормидонтова

Адрес редакции

14-я линия В.О., д. 39, лит. А, г. Санкт-Петербург, 199178, Россия

e-mail: ia@spcras.ru, сайт: http://ia.spcras.ru

Журнал индексируется в международной базе данных Scopus

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий,
в которых должны быть опубликованы основные научные результаты диссертации
на соискание ученой степени доктора и кандидата наук»

Журнал выпускается при научно-методическом руководстве Отделения нанотехнологий
и информационных технологий Российской академии наук

CONTENTS

СОДЕРЖАНИЕ

УДК 519.872 **DOI** 10.15622/ia.23.6.1

В.А. Гончаренко, А.Д. Хомоненко, Р. Абу Хасан

КОМПОЗИЦИОННЫЙ ПОДХОД К ИМИТАЦИОННОМУ МОДЕЛИРОВАНИЮ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ СО СЛУЧАЙНЫМИ ПАРАМЕТРАМИ

Гончаренко В.А., Хомоненко А.Д., Абу Хасан Р. **Композиционный подход к имитационному моделированию систем массового обслуживания со случайными параметрами.**

Аннотация. Обоснован общий подход к моделированию случайных процессов обслуживания в условиях возмущений и неопределенности исходных данных. Предложен композиционный подход построения имитационных моделей массового обслуживания с параметрической неопределенностью на основе распределений фазового типа и фазовых функций. Проведены расчет и сравнение характеристик разработанных имитационных моделей с аналитическими решениями для подтверждения их эффективности и точности. Освещена проблематика неопределенности исходных данных и их влияние на моделирование систем обслуживания. Подчеркивается важность учета параметрической неопределенности в имитационных моделях для повышения их адекватности и применимости на практике. Проведенное исследование включает описание общего подхода к моделированию случайных процессов обслуживания с неопределенностью, а также методологические основы применения фазовых распределений и функций в композиционном моделировании. Рассмотрены четыре класса моделей обслуживания, отличающихся типом интегрального ядра и фазовой функции, что позволяет реализовать разнообразие случайных процессов обслуживания с учетом их особенностей и условий их возникновения. Проведен анализ модели с экспоненциальным интегральным ядром и различными видами фазовых функций, что демонстрирует гибкость и широкие возможности предложенного композиционного подхода к изучению и моделированию систем обслуживания. Представлены результаты имитационного моделирования, подтверждающие аналитические исследования и показывающие применимость и эффективность разработанного подхода при построении и анализе моделей систем обслуживания со случайными параметрами. Отмечается практическая значимость композиционного метода для проектирования и модернизации информационно-вычислительных систем на различных этапах их развития с учетом неопределенности исходных данных. Рассмотрены примеры расчета характеристик элементов архитектуры АСУ железнодорожного транспорта обработки информации в среде имитационного моделирования GPSS World для сетевого узла и сетевой модели массового обслуживания сегмента. Работа ориентирована на развитие методов имитационного моделирования систем массового обслуживания и открывает новые перспективы для их исследования и оптимизации в условиях неопределенности исходных параметров.

Ключевые слова: композиционный подход, интегральное ядро, имитационное моделирование, случайный параметр, параметрическая неопределенность, гипердельтное распределение вероятностей, обобщенная функция, распределение фазового типа, равномерно-экспоненциальное распределение, аппроксимация, фазовая функция, системы массового обслуживания.

1. Введение. Модели систем и сетей массового обслуживания являются важным инструментом для анализа и оптимизации

процессов обслуживания в различных сферах, включая транспорт, телекоммуникации и другие сферы. Эти модели позволяют исследователям и инженерам прогнозировать поведение и характеристики системы, анализировать ее производительность и идентифицировать узкие места, предлагать меры для повышения эффективности работы системы в целом.

В числе актуальных современных направлений развития моделей и методов исследований систем и сетей массового обслуживания можно выделить следующие классы.

1. Модели систем массового обслуживания с управляемыми режимами, включая механизмы "разогрева" и "охлаждения", представляют собой развитие классических подходов к управлению очередями. Эти механизмы позволяют динамически адаптироваться к изменяющимся условиям потока заявок, оптимизируя ресурсы и улучшая качество обслуживания. Для систем облачных вычислений с Веб-интерфейсом в статье [1] предложены вероятностные модели многоканальных систем с «разогревом», «охлаждением» и аппроксимирующими распределениями фазового типа. Работа многих реальных систем, например, серверов дата-центров, сопровождается нагревом и охлаждением сервера. В статье [2] рассматривается задача оптимального выбора пороговых значений разогрева и охлаждения сервера в соответствии с выбранным экономическим критерием.

2. Во многих реальных системах очередей задачи или клиенты имеют разные уровни приоритета. Это требует разработки и анализа моделей систем массового обслуживания с приоритетами, чтобы эффективно управлять потоком клиентов в соответствии с их важностью. Расчет приоритетных режимов для многоканальных систем массового обслуживания относится к классу трудно решаемых задач. В частности, в работе [3] рассматриваются усложненные варианты систем массового обслуживания с динамическими приоритетами. В числе методов расчета вероятностно-временных характеристик многоканальных систем массового обслуживания (СМО) с приоритетами отметим предложенный в [4] и развиваемый в [5] подход к разрешению проблемы снижения размерности, основанный на использовании периода непрерывной занятости системы обслуживанием заявок высшего приоритета.

3. Модели систем массового обслуживания с повторными вызовами – это категория моделей, используемая для исследования систем, в которых заявки, не получившие обслуживание с первой попытки, возвращаются в систему для повторной попытки обслуживания. Они широко применяются для анализа разнообразных

систем и процессов, таких как телекоммуникационные сети, системы технической поддержки, колл-центры и многие другие. Их использование обеспечивает более реалистичное моделирование, в частности, систем с повторными вызовами и конфликтами заявок [6]. В работе [7] исследуется эффект эгоистичного обучения в системе очередей, где соревнуются за ресурсы, но раунды не являются полностью независимыми: число пакетов, которые нужно маршрутизировать в каждом раунде, зависит от успеха маршрутизаторов в предыдущих раундах. В работе [8] исследуется усложненный класс моделей систем с приоритетным обслуживанием и повторными заявками.

4. Модели и методы исследования систем массового облуживания с неопределенностью параметров относятся к сравнительно новому направлению. При моделировании процессов обслуживания в информационно-вычислительных системах (ИВС) обычно параметры модели задаются в виде констант на основе исходных данных, получаемых в результате наблюдений за реальной системой либо экспертного оценивания для проектируемой системы. Однако исходным данным для модели зачастую свойственна некоторая степень неопределенности, обусловленная различными причинами. Так, реальные системы обработки информации функционируют в условиях нестабильности основных параметров, воздействия различных возмущающих факторов, приводящих не просто к увеличению разброса случайных показателей качества, но и к их смещению [9]. Кроме того, неточность получаемых исходных данных о наблюдаемой системе влечет неточность задания параметров модели и результатов моделирования [10]. Для гипотетических объектов и процессов неопределенность исходных данных еще более существенна, так как сведения о моделируемой системе весьма ограничены и приблизительны.

В этом случае целесообразно описание такой неопределенности вводить в модель, как составной ее элемент [11, 12]. В работе [13] представлены исследования неопределенности входных данных при стохастическом моделировании, классификация основных направлений исследований и с акцентом на разработках в последние годы. Рассматриваются прикладные исследования, в которых анализируют представления неопределенности входных данных при моделировании реальных стохастических систем в различных отраслях. Отметим, что аналогичные задачи при стохастическом моделировании управления трафиком воздушным движением решаются в статье [14].

В статье [15], при условии отсутствия предварительной информации о входных моделях и средней поверхности отклика системы, предлагается байесовская непараметрическая структура для количественной оценки воздействия обоих источников неопределенности. Для моделирования смеси разнородных распределений используются непараметрические входные модели на основе процесса Дирихле (DPM), которые могут точно отражать важные характеристики реальных данных, такие как мультимодальность и асимметрия.

Существует ряд аналитических решений для систем массового обслуживания с неопределенностью или возмущениями исходных данных [16 – 20]. Однако задачи, возникающие при исследовании сложных систем в условиях неопределенности исходных данных трудно, а подчас и невозможно решить аналитическими методами, поэтому целесообразно прибегать к методам имитационного моделирования [11, 21].

В данной статье предложен композиционный подход к имитационному моделированию систем с неопределенностью, для которых аналитические решения не всегда возможны, а также сравнительный анализ результатов имитационного моделирования с известными аналитическими решениями.

Научная новизна предлагаемого исследования по сравнению с ранее опубликованными работами состоит в следующем:

1) впервые предложены более полные критерии классификации моделей обслуживания с композиционным представлением исходных распределений, предложенные в [13];

2) развит подход гиперпредставления распределений случайных величин, предложенный в [26], введено гиперравномерное распределение, полученное как композиция равномерного интегрального ядра и гипердельтной фазовой функции;

3) предложена идея имитационного моделирования систем с параметрической неопределенностью на основе композиционного подхода, предполагающего двухэтапную генерацию случайных интервалов времени между событиями (генерация случайных параметров – генерация случайных интервалов);

4) предложена композиционная конструкция операторов (блоков) языка GPSS Generate и Advance для генерации случайных чисел, распределенных по законам со случайными параметрами;

5) предложен ряд имитационных моделей, в том числе сетевых, экспериментально подтверждающих гипотезу о смещении средних времен пребывания в системе при рандомизации

параметров [14], что ведет к повышению требований к проектируемым системам обработки информации.

2. Общий подход к моделированию случайных процессов с неопределенностью параметров распределений. Неопределенность описания случайных процессов из-за недостоверности исходных данных и действия возмущающих факторов часто может быть сведена к параметрической неопределенности, которую необходимо в процессе накопления исходных данных либо устранить, либо описать более точно (в случае, если неоднозначность параметров присуща этим случайным процессам изначально). Процедура уменьшения или уточнения этой недостоверности будет зависеть от характера неопределенности, возможностей получения дополнительной информации и корректировки расчетов до принятия решений [12, 16]. Существует ряд методов анализа стохастических систем, в которых исходные данные изменяются случайным образом [22]. Одним из способов описания моделей с неопределенностью параметров является интервальный подход [24, 25], позволяющий описать размытость, нечеткость параметров с помощью задания диапазонов их возможных значений. В подобных случаях используют методы принятия решений в условиях неопределенности и адаптации [26].

Рассмотрим общий подход к моделированию случайных процессов в системах массового обслуживания (СМО) с неопределенностью исходных данных и внешними возмущающими воздействиями с помощью задания случайности параметров основных распределений вероятностей, описывающих СМО. Будем считать, что случайное изменение данных параметров может определяться как динамикой функционирования сложной системы, так и возникновением возмущающих воздействий, приводящих к отклонению параметров распределений от исходных невозмущенных значений [12]:

$$\hat{\lambda}_i(t) = \lambda_i(t) + \Delta\hat{\lambda}_i(t), \quad i = 1 \div m. \tag{1}$$

Такая модель позволяет учесть, как недостоверность исходных параметров, так и их возмущение. Очевидно, что моменты изменения случайных параметров распределений времени между событиями определяются моментами наступления событий. Поскольку параметры распределений сами формально являются случайными, то функция распределения (ФР) времени между событиями представляется

элементарной случайной функцией – $F(x,\hat{\lambda}) = \hat{y}(x)$. Полной характеристикой такой функции является *функционал распределения* $G(y(x)) = P\{\hat{y}(x) < y(x)\}$, а математическим ожиданием – функция распределения случайной величины (СВ) $\hat{x}$, усредненная по случайным параметрам $\hat{\lambda}_i$.

Если исходную случайную плотность распределения $f(t,\lambda)$ заменить усредненной плотностью распределения, это сведет описываемый случайный процесс к *процессу восстановления* [17]. Последний можно считать математическим ожиданием случайного процесса со случайно распределенными параметрами. При значительном увеличении дисперсии аппроксимируемого случайного процесса качество такой аппроксимации будет ухудшаться [21, 27].

Использование данного приема позволило исследовать ряд новых моделей теории очередей с неопределенностью параметров распределений [16]. В таких моделях вводится понятие *фазовой функции* (или *функции фазы*), как обобщение понятия набора фаз (последовательных, параллельных), используемого в методе фиктивных фаз Эрланга [16, 18].

Рассматриваемый композиционный подход состоит в представлении распределений вероятностей в виде двухуровневой композиции интегрального ядра и фазовой функции, являющейся обобщением понятия плотности распределения случайного параметра. Функцию плотности распределения (ПР) СВ t запишем в виде *композиции* интегрального ядра $f(t,\lambda)$ и фазовой функции $h(\lambda)$ [29]:

$$f(t) = \int_{-\infty}^{\infty} f(t,\lambda)h(\lambda)d\lambda. \qquad (2)$$

Аналитические решения для таких СМО были найдены при простейших предположениях об исходных распределениях (экспоненциальные распределения времени между событиями и равномерные распределения случайных параметров) [12, 17, 18]. Однако теоретически интегральное ядро $f(t,\lambda)$ в формуле (2) можно задать произвольной функцией (степенной, тригонометрической, гиперболической, логарифмической, равномерной и др.). Фазовая же функция может быть задана даже *обобщенной* функцией типа дельта-функции Дирака или её производной [28, 29], при этом по физическому смыслу она может и не быть плотностью распределения случайного параметра. Наибольших результатов можно добиться, если

выбрать в качестве интегрального ядра экспоненту. Если вместо фазовой функции использовать линейные комбинации или производные дельта-функций (в том числе так называемое гипердельтное представление), то можно получить семейство известных *распределений фазового типа* [30 –32]. При выборе же фазовых функций, представляющих плотности распределений непрерывных СВ в результате композиции могут быть получены новые смешанные распределения, которые являются усредненными по случайным непрерывным параметрам функциями [29]. Однако общее аналитическое решение уравнения (2) возможно только в частных случаях для $f(t,\lambda)$ и $h(\lambda)$.

3. Классы моделей обслуживания с параметрической неопределенностью. Распределения фазового типа нашли наибольшее применение в теории массового обслуживания и теории надежности ввиду простоты марковизации случайных процессов путем *свертки* или *вероятностной смеси* экспоненциальных распределений. В то же время экспоненциальное распределение можно считать частным случаем распределения фазового типа с одной фазой.

Однако существует и ряд моделей с распределениями нефазового типа – детерминированным, равномерным, нормальным, гамма-распределением, Парето, Вейбулла и др. К этому же классу можно отнести и так называемое гипердельтное распределение [30], образованное смесью дельта-функций Дирака. В имитационном моделировании также используют смеси и свертки дискретных и непрерывных распределений [18].

В результате исследований было выявлено, что ряд известных распределений является сверткой или смесью других известных более простых распределений [9, 12].

В [16] приведены четыре класса моделей СМО, построенных на основе композиционного представления вида (2).

1) с экспоненциальным интегральным ядром и дискретной фазовой функцией;

2) с экспоненциальным интегральным ядром и фазовой функцией, не являющейся ПР;

3) с ядром фазового типа и непрерывной фазовой функцией;

4) с ядром нефазового типа и произвольной фазовой функцией.

Модели *первого класса* являются наиболее простыми и позволяют получить стандартные фазовые распределения (типа гиперэкспоненциального, гипоэкспоненциального, Кокса и др.).

При экспоненциальном интегральном ядре $f(t,\lambda)$ и гипердельтной фазовой функции $h(\lambda)$

$$h(\lambda) = \sum_{i=1}^{n} C_i \delta(\lambda - \lambda_i) \tag{3}$$

из (2) получим *гиперэкспоненциальную плотность распределения*, часто используемую для аппроксимации реальных распределений случайных величин в теории очередей [34]:

$$f(t) = \sum_{i=1}^{n} C_i \lambda_i e^{-\lambda_i t} . \tag{4}$$

Аналогично можно получить композиционное представление *гипоэкспоненциальной плотности распределения* с коэффициентом вариации меньше 1, если использовать так называемую *гиподельтную фазовую функцию h(λ):*

$$h(\lambda) = \sum_{i=1}^{n} (-1)^{i-1} C_i \delta(\lambda - \lambda_i), \tag{5}$$

$$f(t) = \sum_{i=1}^{n} (-1)^{i-1} C_i \lambda_i e^{-\lambda_i t} . \tag{6}$$

Модели *второго класса* также могут служить для построения фазовых распределений (в частности, последовательных фазовых распределений типа Эрланга). Например, при экспоненциальном интегральном ядре и фазовой функции, представленной частной производной δ-функции Дирака по переменной λ

$$h(\lambda) = \frac{\lambda_0^{\,2} \cdot d\delta(\lambda - \lambda_0)}{\lambda d\lambda}, \tag{7}$$

получим простое распределение Эрланга 2-го порядка

$$f(t) = \lambda_0^2 t e^{-\lambda_0 t} . \tag{8}$$

Модели *третьего класса* в качестве ядра могут использовать фазовые распределения (экспоненциальное, гиперэкспоненциальное, гипоэкспоненциальное, Эрланга, Кокса и др.), а с помощью непрерывной фазовой функции создавать размытые по случайным (возмущенным) параметрам распределения. Например, при экспоненциальном интегральном ядре и непрерывном равномерном представлении фазовой функции

$$h(\lambda) = \frac{\mathbf{1}(\lambda - a) \cdot \mathbf{1}(b - \lambda)}{b - a}, \tag{9}$$

где $\mathbf{1}(x)$ – единичная импульсная функция, образуем *равномерно-экспоненциальное распределение*:

$$f(t) = \frac{(1 + at)e^{-at} - (1 + bt)e^{-bt}}{(b - a)t^2}. \tag{10}$$

Аналогично может быть получено *равномерно-гиперэкспоненциальное распределение* при гиперэкспоненциальном ядре и равномерных распределениях его параметров $\hat{\lambda}_i$. Гиперэкспоненциальное распределение обычно используется для моделирования времен обработки, которые имеют большую вариативность, чем стандартное экспоненциальное распределение. В случае введения равномерного распределения его параметров можно обеспечить еще большую гибкость и дополнительные возможности аппроксимации различных типов данных.

Следует отметить, что аналогичное распределение может быть получено и при экспоненциальном ядре и так называемом *гиперравномерном* представлении фазовой функции. Гиперравномерное распределение представляет собой смесь равномерных распределений случайного параметра. Полученная плотность распределения в соответствии с (2) представляет собой *гиперравномерно-экспоненциальное распределение*. Как показали проведенные исследования, несмотря на различные схемы образования этих распределений, итоговые формулы совпадают. Оба подхода в конечном счёте приводят к интегрированию экспоненциальной функции по случайному параметру $\hat{\lambda}$, распределённому равномерно на нескольких интервалах. В случае гиперравномерно-экспоненциального распределения это происходит

через смесь равномерных распределений случайного параметра, а в случае равномерно-гиперэкспоненциального распределения – через прямое равномерное распределение этих параметров гиперэкспоненты.

При экспоненциальном ядре и экспоненциальной фазовой функции усредненное по случайному параметру распределение представляет собой смещенное распределение Парето 1-го порядка, у которого математическое ожидание бесконечно. Можно показать, что формирование распределений Парето высших порядков на основе композиционного подхода потребует использования новых операторов преобразования, например, таких, как степенные или логарифмические функции, или случайности высших степеней (например, случайность параметров распределения случайного параметра $\hat{\lambda}$).

Наиболее общие модели *четвертого класса* позволяют создавать сколь угодно сложные распределения. К простейшему распределению этого класса относится детерминированное (регулярное) распределение, у которого и интегральное ядро, и фазовая функция представлены дельта-функциями Дирака.

При детерминированном интегральном ядре, представленном дельта-функцией Дирака $\delta(t-T)$, и гипердельтном представлении фазовой функции

$$h(T)=\sum_{i=1}^{n} C_i \delta(\mathrm{T}-\mathrm{T}_i) \tag{11}$$

имеем гипердельтную плотность распределения

$$f(t)=\sum_{i=1}^{n} C_i \delta(t-T_i). \tag{12}$$

При непрерывном равномерном представлении интегрального ядра плотность распределения имеет вид

$$f(t,T)=\frac{\mathbf{1}(t-u)\cdot\mathbf{1}(\nu-t)}{\nu-u}, \tag{13}$$

где $T = (u+v)/2, u = T - \sigma\sqrt{3}, u = T + \sigma\sqrt{3}$, и при гипердельтной фазовой функции получим *гиперравномерное представление* плотности распределения

$$f(t) = \sum_{i=1}^{n} C_i \frac{\mathbf{1}(t-u_i)\cdot\mathbf{1}(v_i-t)}{v_i-u_i}. \qquad (14)$$

Выделим критерии классификации моделей обслуживания с композиционным представлением исходных распределений:

– *по типу интегрального ядра:*

1) с ядром фазового типа (включая экспоненциальное);

2) с ядром нефазового типа;

– *по физической интерпретации фазовой функции:*

1) с фазовой функцией, являющейся плотностью распределения параметра;

2) с фазовой функцией, не являющейся плотностью распределения параметра;

– *по типу фазовой функции:*

1) с фазовой функцией, являющейся непрерывным распределением;

2) с фазовой функцией, являющейся дискретным распределением;

3) с фазовой функцией, являющейся смешанным распределением.

Выделенные классы распределений удобно представить в виде таблицы 1.

Таблица 1. Классы распределений с композиционным представлением

Тип фазовой функции	Физическая интерпретация фазовой функции	Интегральное ядро	
		Ядро фазового типа	Ядро нефазового типа
Дискретная фазовая функция (discrete)	плотность распределения	1Dd	2Dd
	не является плотностью	1Dn	2Dn
Непрерывная фазовая функция (continuos)	плотность распределения	1Cd	2Cd
	не является плотностью	1Cn	2Cn
Смешанная фазовая функция (mixed)	плотность распределения	1Md	2Md
	не является плотностью	1Mn	2Mn

Однако для многих из этих моделей аналитические решения не всегда возможны, поэтому предложенный композиционный подход [29] можно применить при имитационном моделировании

систем [21, 27, 35]. Предварительно для простых моделей необходимо провести сравнительный анализ полученных результатов имитационного моделирования с известными аналитическими решениями.

4. Построение имитационных моделей с параметрической неопределенностью на основе композиционного подхода. Сформулируем постановку задачи имитационного моделирования. Дана одноканальная система массового обслуживания, для которой заданы: функция распределения (ФР) интервалов времени между входными заявками *A(t),* ФР времени обслуживания заявок *B(x)* и функции распределения $h_1(\lambda_i)$, $h_2(\mu_i)$ случайных параметров $\hat{\lambda}_i$, $\hat{\mu}_i$ распределений *A(t)* и *B(x)*. Необходимо построить имитационную модель, позволяющую исследовать процессы обслуживания при различных коэффициентах вариации и степенях неопределенности параметров распределений.

В качестве интегральных ядер будем использовать различные распределения фазового типа. Простейшей моделью с возмущениями является система $\hat{M}/\hat{M}/1$ с экспоненциальными *A(t)* и *B(x)* при случайных параметрах $\hat{\lambda}$ и $\hat{\mu}$ соответственно.

Также для моделирования входящих потоков и потоков обслуживания с коэффициентами вариации $\upsilon>1$ используем гиперэкспоненциальные распределения с ФР вида:

$$A(t) = 1 - C_1 e^{-\lambda_1 t} - (1 - C_1) e^{-\lambda_2 t} \tag{15}$$

или обобщенные распределения Эрланга (при $\upsilon<1$) с ФР:

$$A(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \left(e^{-\lambda_1 t} - e^{-\lambda_2 t} \right). \tag{16}$$

Выражения для *B(x)* имеют аналогичный вид.

Формирование интервалов между событиями исходных распределений в модели будем производить в два этапа: на первом формируются случайные значения параметров распределений, а на втором – непосредственно сам интервал в соответствии с выбранными значениями параметров. Это позволяет более адекватно воспроизводить исследуемые процессы, не усредняя их по случайным параметрам, поскольку в соответствии с [12] усредненные по параметру

распределения являются математическими ожиданиями случайных функций.

Для генерации равномерных законов распределений случайных параметров $h_1(\lambda_i)$, $h_2(\mu_j)$, как и для экспоненциальных законов распределений *A(t)* и *B(x),* можно использовать метод обратной функции [21]. Например, для случайных параметров $\hat{\lambda}_i$, равномерно распределенных в диапазоне $[a_i, b_i]$, получим:

$$\hat{\lambda}_i = (b_i - a_i)\hat{\xi}_i + a_i, \tag{17}$$

где $\hat{\xi}_i$ – *i*-я случайная величина, распределенная по равномерному закону в диапазоне [0,1]. Аналогично формируются значения $\hat{\mu}_j$.

Генерация экспоненциально распределенной случайной величины $\hat{X}_k$ со случайным параметром $\hat{\lambda}_i$ осуществляется на основе формулы

$$\hat{X}_k = -(1/\hat{\lambda}_i)\ln(1-\hat{U}_k), \tag{18}$$

где $\hat{\lambda}_i$ берется из ранее полученных вычислений (17).

Специфика генераторов случайных чисел для имитационного моделирования фазовых распределений типа *гиперэкспоненциального распределения* и *распределения Эрланга* состоит в том, что они также выполняются из двух подэтапов.

Обобщенное распределение Эрланга r-го порядка со случайными параметрами предполагает последовательное прохождение *r* фаз и соответственно является распределением суммы *r* экспоненциально распределенных чисел, у каждого из которых свой параметр $\hat{\lambda}_i$, случайно выбираемый из соответствующего диапазона $[a_i, b_i]$. Следовательно, из (18) имеем

$$\hat{X} = \sum_{i=1}^{r}\left(-\frac{1}{\hat{\lambda}_i}\ln(1-\hat{U}_i)\right). \tag{19}$$

Формируемый случайный поток можно интерпретировать двумя способами:

1) как возмущенный вариант исходного обобщенного потока Эрланга с неслучайными параметрами $\overline{\lambda}_i$, где $\hat{\lambda}_i = \overline{\lambda}_i + \hat{\eta}_i$, $\hat{\eta}_i \in \left[-(b_i - a_i)/2; (b_i - a_i)/2\right]$.

2) как рандомизированный обобщенный поток Эрланга, имеющий более широкий диапазон изменения коэффициента вариации, чем у исходного распределения. Например, начальные моменты такого распределения 2-го порядка могут быть получены интегрированием условных начальных моментов по всем возможным значениям $\hat{\lambda}_i$:

$$\begin{aligned} \overline{\nu}_1 &= \frac{\ln(b_1/a_1)}{b_1 - a_1} + \frac{\ln(b_2/a_2)}{b_2 - a_2}; \\ \overline{\nu}_2 &= \frac{2}{a_1 b_1} + 2\frac{\ln(b_1/a_1)}{b_1 - a_1} \cdot \frac{\ln(b_2/a_2)}{b_2 - a_2} + \frac{2}{a_2 b_2}. \end{aligned} \qquad (20)$$

Гиперэкспоненциальное распределение r-го порядка предполагает параллельное прохождение фаз, но также требует двух обращений. При первом обращении выбирается номер фазы $1 < i < r$, а при втором обращении генерируется показательное распределение с выбранным параметром $\hat{\lambda}_i$ из *i*-й фазы.

Разработанная программа имитационного моделирования, написанная на языке *Python*, позволяет моделировать системы массового обслуживания как со случайными, так и неслучайными параметрами распределений, а также использовать различные типы распределений времени между заявками и времени обслуживания в узле: экспоненциальное, гиперэкспоненциальное и обобщенное эрланговское. Исходными данными для моделирования являются:

– тип аппроксимации исходных распределений (по коэффициенту вариации для входного потока заявок υ_1 и распределения времени обслуживания υ_2);

– границы параметров [*a, b*] и [*c, d*] соответствующих распределений;

– предел моделирования в виде числа обслуженных заявок или времени моделирования.

В результате моделирования определяются следующие характеристики:

– начальные моменты распределений времени между заявками и времени обслуживания α_j и β_j;

– коэффициент загрузки системы ρ;

– начальные моменты распределения времени ответа системы γ_j.

Результаты имитационного моделирования СМО типа $\hat{M}/\hat{M}/1$ со случайными параметрами, распределенными по равномерным законам, представлены в таблице 2. При этом проводилась обработка 10000 заявок в каждом случае при холостом прогоне 500 заявок для вхождения системы в установившийся режим работы. В основном, результаты имитации подтверждают результаты аналитических исследований, проведенных в [16, 17], и известные оценки систем массового обслуживания.

Таблица 2. Результаты имитационного моделирования СМО типа $\hat{M}/\hat{M}/1$

a, 1/с	b, 1/с	c, 1/с	d, 1/с	Нач. моменты γ_j	Аналит. результат	Имитац. результат	Погрешность, %
1,0	1,0	2,0	2,0	γ_1, с γ_2, с^2 γ_3, с^3	1,0 2,0 6,0	0,989 1,922 5,690	1,11 4,12 5,45
1,0	2,0	3,0	4,0	γ_1, с γ_2, с^2 γ_3, с^3	0,501 0,506 0,768	0,497 0,489 0,686	0,81 3,48 11,95
1,0	5,0	3,0	8,0	γ_1, с γ_2, с^2 γ_3, с^3	0,441 0,412 0,589	0,482 0,453 0,651	-8,5 -9,05 -9,52
1,0	8,0	2,0	10,0	γ_1, с γ_2, с^2 γ_3, с^3	0,901 1,784 5,291	0,852 1,627 4,61	5,63 7,78 13,03

Также оказалось, что время ответа в СМО типа $M/M/1$ больше, чем в системах $E_n/M/1$ и $M/E_n/1$ и меньше, чем в системах $H_n/M/1$ и $M/H_n/1$ как при фиксированных, так и при случайных параметрах (при одинаковой степени неопределенности последних в различных системах). Для систем одного типа случайность параметров увеличивает среднее время ответа и тем больше, чем шире диапазон их изменения, а, следовательно, дисперсия и коэффициент вариации.

Относительно точности имитационного моделирования в сравнении с аналитическими расчетами нужно отметить, что она во многом зависит от исходных данных и от применяемой процедуры генерации псевдослучайных чисел. В основном, для первого момента времени пребывания относительная погрешность не превышает 5-7%, для высших моментов она увеличивается. Кроме того, погрешность моделирования уменьшается при увеличении коэффициента загрузки,

однако, при стремлении коэффициента загрузки к 1 погрешность снова растет, что обусловлено погрешностью имитационного моделирования высоконагруженных систем.

Для общности представления в таблице 3 приведено качественное сравнение преимуществ и недостатков композиционного подхода с численными и аналитическими методами моделирования СМО при неопределенностях.

Таблица 3. Сравнение подходов к моделированию СМО

Критерий / Подход	Композиционный подход	Численные методы	Аналитический расчет
Преимущества	Гибкость при моделировании неопределенности. Модульность и адаптация к изменениям	Подходит для нелинейных систем и оптимизации. Гибкость в моделировании неопределенностей	Высокая точность для простых систем. Не требует вычислительных ресурсов
Недостатки	Зависимость от вычислительных ресурсов. Сложность точной калибровки	Требует значительных вычислительных мощностей. Сложность интерпретации результатов	Ограниченная применимость при неопределенностях. Упрощение реальности и ограничения в моделях
Управление неопределен-ностями	Возможность легкой интеграции различных сценариев	Может аппроксимировать поведение систем с неопределенными параметрами	Проблематично при сильных неопределенностях параметров
Зависимость от ресурсов	Высокая (особенно для сложных моделей)	Высокая (для вычислительно сложных задач)	Низкая (возможность аналитических расчетов вручную)
Применимость моделей	Любая степень сложности и неопределенности	Лучше подходит для сложных систем и оптимизационных задач	Преимущественно для простых и хорошо изученных систем
Интерпретация результатов	Может быть сложной из-за объема и сложности данных	Сложна без специальных знаний	Обычно прямолинейна для изученных моделей

5. Пример реализации композиционного подхода. Рассмотрим пример реализации расчета характеристик элементов архитектуры автоматизированной системы управления железнодорожного транспорта (АСУ ЖТ) обработки информации в среде имитационного моделирования GPSS World (свободно распространяемая система, предназначена для имитационного моделирования сложных систем с дискретными и непрерывными процессами, обладает гибкими возможностями настройки

и визуализации моделей, имеет средства для моделирования систем и сетей массового обслуживания). Сначала рассмотрим отдельный сетевой узел и способы реализации его имитационной модели (рисунок 1).

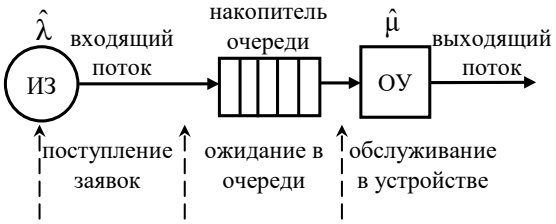


Рис. 1. Модель СМО со случайными параметрами

Для имитации случайности параметров распределений времени между входящими заявками и времени обслуживания была реализована конструкция блоков GPSS с двойной генерацией случайной величины. Например, блок GPSS

Generate (Exponential (1, 0, (1/Uniform (2, a, b))))

позволяет сначала сгенерировать по равномерному закону (Uniform) случайный параметр $\hat{\lambda}$, изменяющийся в диапазоне [a, b], затем найти обратную ему величину случайного математического ожидания времени между входящими заявками, выступающую в качестве параметра в блоке Generate, и потом использовать ее для генерации собственно случайного времени между входящими заявками. При вызове Exponential, Uniform и других функций распределения первыми параметрами являются номера генераторов псевдослучайных чисел от 1 до 999, которые рекомендуется в одной программе для различных случайных величин делать разными, чтобы обеспечить независимость имитируемых величин.

Аналогичный прием используется в блоке Advance для генерации времени обслуживания со случайным параметром $\hat{\mu}$:

Advance (Exponential(3,0,(1/Uniform(4, c, d)))).

Блоки Generate и Advance по формату вызова отличий не имеют, различие в названии блоков обусловлено разными моделируемыми с их помощью случайными величинами – интервалов

между смежными заявками и длительностей обслуживания, соответственно.

При моделировании для удобства учета степени неопределенности случайных параметров распределений можно использовать так называемые коэффициенты неопределенности

$$\Delta_\alpha = (b-a)/(b+a),\ \Delta_\beta = (d-c)/(d+c),$$

принимающие значения от 0 до 1. Отсутствию неопределенности соответствуют нулевые значения коэффициентов, при увеличении степени неопределенности параметров коэффициенты стремятся к 1.

Из таблицы 4 видно, что при расширении границ изменения интенсивности обслуживания увеличивается среднее время обслуживания, и как следствие, растет коэффициент загрузки и среднее время ожидания. При увеличении же диапазона изменения интенсивности входящего потока загрузка и время ожидания падают. При одновременном увеличении неопределенности обоих параметров загрузка остается примерно одинаковой, но время ожидания растет.

Таблица 4. Имитационное моделирование систем с неопределенностью

Модель	Границы [a,b], Δ_α	Границы [c,d], Δ_β	Среднее время обслуживания	Загрузка	Среднее время ожидания	Средняя длина очереди
*M/M/*1	8, 0	10	0,100	0,799	0,307	2,447
$M / \hat{M}$ /1	8, 0	[9,11], 0,1	0,101	0.802	0,313	2,495
	8, 0	[8,12], 0,2	0,102	0,810	0,333	2,655
	8, 0	[5,15], 0,5	0,110	0,878	0,614	4,894
	8, 0	[4,16], 0,6	0,116	0,923	1,122	8,946
	8, 0	[2,18], 0,8	0,138	1,00	584,201	4658,51
	8, 0	[1,19], 0,9	0,164	1,00	1907.79	15206.72
$\hat{M} / M$ /1	[7,9], 0,125	10, 0	0,100	0,795	0,303	2,407
	[6,10], 0,25	10, 0	0,100	0,782	0,291	2,271
	[4,12], 0,5	10, 0	0,100	0,728	0,246	1,786
	[3,13], 0,625	10, 0	0,100	0,682	0,219	1,491
	[2,14], 0,75	10, 0	0,100	0,617	0,191	1,178
	[1,15], 0,875	10, 0	0,100	0,517	0,164	0,845
$\hat{M} / \hat{M}$ /1	[7,9], 0,125	[9,11], 0,1	0,101	0,798	0,309	2,454
	[6,10], 0,25	[8,12], 0,2	0,102	0,793	0,314	2,453
	[4,12], 0,5	[5,15], 0,5	0,110	0,8	0,416	3,020
	[3,13], 0,625	[4,16], 0,6	0,116	0,788	0,467	3,177
	[2,14], 0,75	[2,18], 0,8	0,138	0,848	1,119	6,889
	[1,15], 0,875	[1,19], 0,9	0,164	0,849	2,088	10,779

В таблице 5 приведены результаты оценивания доверительных интервалов $\Delta\gamma_1^{\text{имит}}$ среднего времени пребывания в системе $\gamma_1^{\text{имит}}$ при имитационном моделировании в зависимости от числа испытаний при различных видах неопределенности (параметров входного потока и обслуживания). Рассмотрены 4 случая, когда случайные интенсивности входящего потока $\hat{\lambda}$ и потока обслуживания $\hat{\mu}$ либо усредняются (0.05 и 0.1), либо рассматриваются равномерно распределенными в заданных диапазонах (0.02-0.08 и 0.04-0.16 соответственно). Приведены также данные аналитического расчета $\gamma_1^{\text{аналит}}$.

Таблица 5. Оценка доверительных интервалов среднего времени пребывания в системе по результатам имитационного моделирования при $p_{\text{дов}} = 0{,}95$

	$\hat{\lambda}$	$\hat{\mu}$	$\gamma_1^{\text{аналит}}$	N испыт	$\gamma_1^{\text{имит}}$	$\Delta\gamma_1^{\text{имит}}$	$\gamma_1 - \Delta\gamma_1$	$\gamma_1 + \Delta\gamma$
M/*M*/1	0,05	0,1	20	500	20,240	1,782	18,458	22,022
				1000	21,816	1,407	20,409	23,223
				10000	19.726	0,369	19,357	20,095
$M/\hat{M}/1$	0,05	0,04-0,16	30,049	500	31,845	2,891	28,894	34,736
				1000	35,513	2,530	32,983	38,043
				10000	29,870	0,618	29,252	30,488
$\hat{M}/M/1$	0,02 – 0,08	0,1	18,874	500	20,163	1,752	18,411	21,915
				1000	20,371	1,259	19,112	21,630
				10000	18,733	0,356	18,377	19,089
$\hat{M}/\hat{M}/1$	0,02 – 0,08	0,04-0,16	27,130	500	30,846	2,785	28,061	33,631
				1000	30,669	2,142	28,527	32,811
				10000	27,028	0,565	26,463	27,593

Результаты исследований подтверждают выводы [13 – 15], что даже если при моделировании затрачивается большой объем вычислительных усилий для улучшения оценки показателя производительности, его оценка будет подвержена значительной изменчивости из-за неопределенности в точечных оценках параметров распределений. Если неточность входных данных не учитывается должным образом, то дополнительные усилия по моделированию могут привести к еще более низким доверительным интервалам с меньшим охватом. При малом числе испытаний достаточно широкий доверительный интервал перекрывает погрешности переходного периода. При увеличении же числа испытаний сужается доверительный интервал, и, несмотря на увеличение точности моделирования, есть риск исключения истинного значения параметра из доверительного интервала. В [15] используется непараметрический подход при неопределенности для улучшения оценки доверительных интервалов.

Очевидно, что сравнение систем удобнее проводить при одинаковых коэффициентах загрузки, которые меняются при изменении коэффициентов неопределенности из-за смещения средних времен между поступлениями и обслуживания.

При рассмотрении зависимости относительного среднего времени ожидания (к среднему времени обслуживания) ω_1/β_1 от коэффициентов неопределенности Δ_α и Δ_β при фиксированном коэффициенте загрузки ρ=0,5 (рисунок 2) можно увидеть, что характеристики СМО с неопределенностью более чувствительны к случайности параметров обслуживания, чем к случайности параметров входного потока [16].

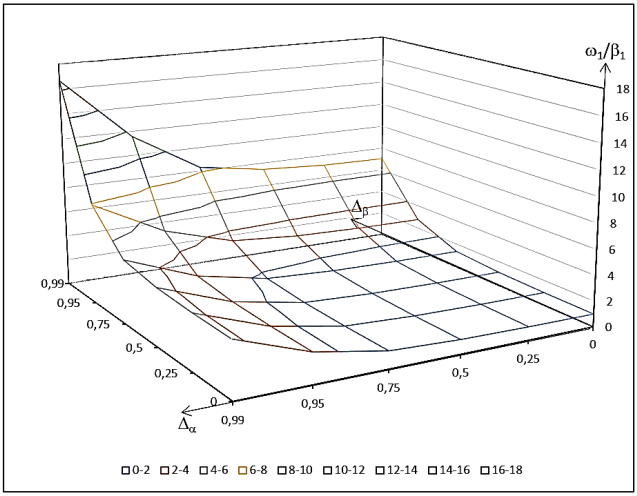


Рис. 2. Зависимость ω_1/β_1 от коэффициентов Δ_α и Δ_β

Генерацию случайных чисел, подчиняющихся распределениям фазового типа, отсутствующим в библиотечных ГСЧ GPSS World, например, распределения Эрланга, можно выполнить с помощью его обобщенной формы – гамма-распределения, при целом положительном параметре *a*, задающем форму распределения (число фаз).

Математическое ожидание времени обслуживания в этом случае будет определяться по формуле: M_β=ab+s, где s=0 – смещение, b – среднее время обслуживания на одной фазе. Для реализации распределения Эрланга 2-го порядка со случайным параметром b, распределенным в диапазоне [c, d], применим выше описанную схему:

Advance (Gamma(6,0,(1/Uniform(7, c, d)),2)).

Кроме того, для генерации случайных времен со случайными параметрами может использоваться схема с тройной генерацией:

1) генерируются случайные реализации случайных параметров распределений фазового типа;

2) генерируются случайные времена задержки на каждой из фиктивных фаз в соответствии с полученными параметрами;

3) выполняется процедура аддитивного или вероятностного смешивания случайных задержек.

Например, для реализации генерации времени обслуживания по закону Эрланга 2-го порядка со случайными параметрами $\hat{\mu}_i$ используется аддитивная схема (суммирование задержек):

Advance(Exponential(5,0,(1/Uniform(6,c1,d1)))+Exponential(7,0,(1/Uniform(8,c2,d2)))).

При различных границах диапазонов изменения параметров [c1,d1], [c2,d2] может быть реализовано обобщенное распределение Эрланга со случайными параметрами. При одинаковых границах реализуется простое распределение Эрланга.

Для генерации гиперэкспоненциального распределения используется схема с логическими переключателями [36], имеющими два состояния (включен, выключен), поскольку для генерации случайной величины нужно будет выбрать только одну из случайных величин с заданной вероятностью (листинг 1).

```
*****************************Область описания*********************************
Ver1        EQU     0.396       ; Вероятность перехода транзакта на Kluch1
Ver2        EQU     (1-Ver1)    ; Вероятность перехода транзакта на Kluch2
T1          EQU     26.33       ; Время обслуживания транзакта в 1 фазе
T2          EQU     15.85       ; Время обслуживания транзакта во 2 фазе
************************Имитация работы ключей*********************************
GENERATE ,,,1           ; Формирование 1 транзакта
Variator TRANSFER Ver1,Kl1,Kl2 ; Переход с вероятностью Ver1 в Kl2
Kl1         LOGIC S     Kluch1      ; Включение ключа Kluch1
ADVANCE (T1#5)          ; Задержка транзакта в блоке на T1*5 ед. модельного времени
            LOGIC R     Kluch1      ; Выключение ключа Kluch1
            TRANSFER Variator ; Безусловный переход переключатель ключей
Kl2         LOGIC S     Kluch2      ; Включение ключа Kluch2
ADVANCE (T2#5)          ; Задержка транзакта в блоке на T2*5 ед. модельного времени
            LOGIC R     Kluch2      ; Выключение ключа Kluch2
            TRANSFER ,Variator; Безусловный переход в переключатель ключей
```

Листинг 1. Имитация работы ключей

Код имитации обслуживания заявок по гиперэкспоненциальному закону 2-го порядка с помощью блока GATE приведен на листинге 2.

```
**********Имитация обслуживания заявок**********
GENERATE (Exponential(11,0,10)) ; Формирование простейшего потока
GATE LS    Kluch1,Met10 ; Если ключ Kluch1 включен, то переход к следующему
                        ; блоку, иначе переход на Met10
SEIZE Uzel ; Попытка занять устройство Uzel
ADVANCE (Exponential(1,0,T1)) ; Обслуживание заявки в 1 фазе
RELEASE Uzel ; Освобождение устройства Uzel
TRANSFER ,Met20 ; Безусловная передача транзакта на выход из системы
Met10      SEIZE      Uzel  ; Попытка занять устройство Uzel
ADVANCE (Exponential(2,0,T2)) ; Обслуживание заявки во 2 фазе
RELEASE Uzel ; Освобождение устройства Uzel
Met20      TERMINATE 1
```

Листинг 2. Имитация обслуживания заявок

Теперь рассмотрим пример реализации сетевой модели массового обслуживания сегмента АСУ ЖТ из трех обслуживающих узлов (0-й узел – исток, 4-й узел – сток), приведенной на рисунке 3.

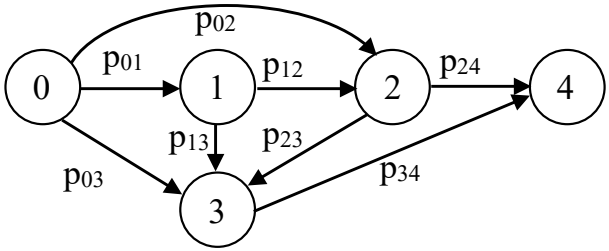


Рис. 3. Сетевая модель обслуживания сегмента АСУ ЖТ

Дано: узлы одноканальные, дисциплина обслуживания FIFO, очереди к узлам не ограничены. Пусть узел 1 реализует *Модуль управления ресурсами станций*, на который поступают данные о времени прибытия/отправления поездов. Узел 2 реализует *Базу данных с интеграционным слоем*, которая обеспечивает хранение данных и выдачу результатов запросов. На него поступают актуальные данные о состоянии ресурсов станций с узла 1, а также данные от внешних узлов. 3 узел имитирует *Модуль прогнозирования и аналитики*, на который поступают оперативные и исторические данные для анализа. Результатом обработки на узле 3 являются аналитические отчеты и рекомендации для оптимизации работы.

Нулевой узел имитирует источник заявок сети (в частности, от *Модуля управления расписанием*, *Модуля управления движением* и т.д.), узел 4 – сток (выход сети, результаты обработки данных).

Примем следующие вероятности передач между узлами сети: $p_{01} = 0{,}5$; $p_{02} = 0{,}4$; $p_{03} = 0{,}1$; $p_{12} = 0{,}7$; $p_{13} = 0{,}3$; $p_{23} = 0{,}4$; $p_{24} = 0{,}6$; $p_{34} = 1$. Плотности распределения интервалов между заявками входного потока $a(t)$ и времени обслуживания заявок в узлах $b_i(x)$ будем считать экспоненциальными, со случайными равномерно распределенными параметрами λ и μ_i: $\lambda \in [a, b]$, $\mu_i \in [c_i, d_i]$. Следует найти характеристики времени обработки заявок в сети без учета и с учетом факторов неопределенности при $\lambda = 10 \pm 8$ мин$^{-1}$, $\mu_1 = 5 \pm 3$ мин$^{-1}$, $\mu_2 = 8 \pm 6$ мин$^{-1}$, $\mu_3 = 10 \pm 6$ мин$^{-1}$.

Если факторы случайности (неопределенности) параметров распределений не учитывать, сеть будет рассчитываться как обычная экспоненциальная сеть массового обслуживания. Но если в качестве исходных данных брать усредненные случайные параметры распределений вероятностей, т.е. $\lambda = (a+b)/2$, $\mu_i = (c_i+d_i)/2$, получим искаженные средние времена между входящими заявками α_1 и времена обслуживания в узлах β_{1i} со сдвигом влево (таблица 4) и, соответственно, коэффициенты загрузки узлов сети ρ_i (в данном случае завышенные). При учете факторов неопределенности исходных параметров [16, 17] первые начальные моменты времени между входящими заявками α_1 и времени обслуживания в узлах β_{1i} (таблица 1) будут иметь вид:

$$\alpha_1 = \frac{\ln(b/a)}{b-a}, \quad \beta_{1i} = \frac{\ln(d_i/c_i)}{d_i - c_i}. \tag{21}$$

Средние интенсивности потоков $1/\alpha_1$ и $1/\beta_{1i}$ оказываются меньше средних значений случайных параметров λ, μ_i. Это является следствием теоремы 1 из [14], гласящей, что *математическое ожидание случайной неотрицательной величины* $\hat{x}$ *всегда не меньше обратного значения математического ожидания обратной случайной величины* $1/\hat{x}$.

В результате имитационного моделирования получаем средние времена пребывания заданий в узлах γ_{1i}, затем рассчитываем среднее время пребывания заданий в сети γ_1 (таблица 4).

Таблица 4. Результаты моделирования сегмента АСУ ЖТ

Узел	Без учета неопределенности				С учетом неопределенности			
	α_1	β_1	ρ	γ_1	α_1	β_1	ρ	γ_1
1 узел	0,249	0,2	0,82	1,111	0,335	0,231	0,690	1,073
2 узел	0,2	0,125	0,625	0,333	0,275	0,162	0,590	0,595
3 узел	0,133	0,1	0,75	0,4	0,183	0,116	0,631	0,439
Сеть				**1,078**				**1,216**

Введенные ранее коэффициенты неопределенности параметров для рассматриваемого примера принимают следующие значения: $\Delta_\alpha = 0{,}8$, $\Delta_{\beta_1} = 0{,}6$, $\Delta_{\beta 2} = 0{,}75$, $\Delta_{\beta 3} = 0{,}6$.

Из анализа результатов, приведенных в таблице 4, следует, что при учете неопределенности по сравнению с отсутствием учета неопределенности оценка задержки прохождения заявок в модели сегмента АСУ ЖТ увеличивается на 11,35%. Таким образом, результаты имитационного моделирования показали, что учет возможной неопределенности (случайности, неточности, колебаний, в том числе из-за действия внешних дестабилизирующих факторов) параметров распределений позволяет не только показать, что учет неопределенности важен, но и численно по результатам моделирования (с учетом и без учета неопределенности) обосновать, насколько именно он способствует улучшению оценок задержек в обслуживании и формулированию адекватных требований к производительности распределенных вычислительных систем.

6. Заключение. Композиционный подход позволяет проводить как аналитическое, так и имитационное моделирование стохастических систем при наличии неточности, размытости или изменчивости параметров распределений вероятностей, заданных в качестве исходных. При этом предполагается нестандартное использование экспоненциального распределения как основы для создания более сложных распределений, что может привести к новым интересным результатам в стохастических процессах и теории распределений. Но в отличие от аналитических расчетов, имитационное моделирование не ограничивается исследованием простейших экспоненциальных моделей, а позволяет задействовать другие распределения фазового и нефазового типа со случайными параметрами (гиперэкспоненциальное, Эрланга, Кокса, гипердельтное, равномерное и др.).

Сравнивая композиционный подход и численные методы, можно отметить, что оба эти подхода обладают своими уникальными

преимуществами и ограничениями в зависимости от специфики задачи. Однако композиционный подход особенно выделяется своей способностью к анализу систем с неопределенными параметрами благодаря его модульности, гибкости и возможности интегрировать широкий спектр вероятностных распределений. Это делает его особенно ценным на различных этапах проектирования и модернизации ИВС, позволяя более обоснованно предъявлять требования к производительности и надежности ИВС, что критично для поддержания устойчивости их функционирования.

Практическая значимость метода определяется его возможностью использования на различных этапах проектирования и модернизации ИВС в зависимости от имевшегося объема исходных данных, и более обоснованного предъявления требования к производительности ИВС, критичных к нарушению устойчивости функционирования.

Дальнейшие исследования по вопросам практически полезного применения рассматриваемого в статье метода имитационного моделирования, на наш взгляд, целесообразно продолжить в части оценивания доверительных интервалов показателей оперативности систем с неопределенностью [13, 15], анализа возможностей и технологий эффективного использования в современных актуальных сферах обеспечения требуемого уровня оперативности, надежности и кибербезопасности функционирования информационных систем, сетей передачи данных и обработки больших данных [37 – 39].

Литература

1. Khomonenko A., Khalil M., Kassymova D. Probabilistic Models for Evaluating the Performance of Cloud Computing Systems with Web Interface. SPIIRAS Proceedings. 2016. vol. 6. no. 49. pp. 49–65.
2. Dudina O., Dudin A. Optimization of queueing model with server heating and cooling // Mathematics. 2019. vol. 7. no. 9. DOI: org/10.3390/math7090768.
3. Klimenok V., Dudin A., Dudina O., Kochetkova I. Queuing system with two types of customers and dynamic change of a priority // Mathematics. 2020. vol. 8(5). DOI: org/10.3390/math8050824.
4. Khomonenko A.D. Performance analysis of the multiprocessor systems with priority processing of heterogeneous requests flow // Avtomatika i Vychislitel'naya Tekhnika. 1991. no. 4. pp. 55–64.
5. Краснов С.А., Лохвицкий В.А., Хабаров Р.С. Численный анализ многоканальных систем массового обслуживания с абсолютным приоритетом на основе фазовой аппроксимации периода непрерывной занятости // Труды Военно-космической академии имени А.Ф. Можайского. 2022. № 682. С. 7–20.
6. Nazarov A., Strik J., Kvach A. A survey of recent results in finite-source retrial queues with collisions // Information technologies and mathematical modelling. Queueing Theory and Applications: 17th International Conference and 12th

Workshop on Retrial Queues and Related Topics. 2018. vol. 912. pp. 1–15. DOI: 10.1007/978-3-319-97595-5_1.
7. Gaitonde J., Tardos E. Stability and learning in strategic queuing systems // Proceedings of the 21st ACM Conference on Economics and Computation. 2020. pp. 319–347.
8. Заяц О.И., Кореневская М.М., Ильяшенко А.С., Мулюха В.А. Система массового обслуживания с абсолютным приоритетом, вероятностным выталкивающим механизмом и повторными заявками // Информатика и автоматизация. 2024. Т. 23. № 2. С. 325–351.
9. Бусленко Н.П. Моделирование сложных систем. М.: Наука, 1978. 400 с.
10. Ивницкий В.А. Об оценке точности результатов моделирования сложных систем с неточной входной информацией при схеме независимых испытаний // Известия АН СССР. Техническая кибернетика. 1974. Т. 4. С. 208–217.
11. Law A.M. Simulation Modeling and Analysis, 6th Edition. McGraw Hill. 2024. 688 p.
12. Гончаренко В.А. Формальный аппарат представления случайных процессов обслуживания с возмущающими воздействиями и неопределенностью параметров // Труды Военно-космической академии им. А.Ф. Можайского. 2015. № 648. С. 13–18.
13. Corlu C.G., Akcay A., Xie W. Stochastic simulation under input uncertainty: A review // Operations Research Perspectives. 2020. vol. 7. DOI: 10.1016/j.orp.2020.100162.
14. Shone R., Glazebrook K., Zografos K.G. Applications of stochastic modeling in air traffic management: Methods, challenges and opportunities for solving air traffic problems under uncertainty // European Journal of Operational Research. 2021. vol. 292. no. 1. pp. 1–26.
15. Xie W., Li C., Wu Y., Zhang P. A nonparametric Bayesian framework for uncertainty quantification in stochastic simulation // SIAM/ASA Journal on Uncertainty Quantification. 2021. vol. 9. no. 4. pp. 1527–1552.
16. Гончаренко В.А. Модели и методы анализа систем массового обслуживания с параметрической неопределенностью // Интеллектуальные технологии на транспорте. 2017. № 4. С. 5–11.
17. Гончаренко В.А. Моделирование и оценивание характеристик случайных потоков событий в компьютерных сетях при параметрической неопределенности // Труды Военно-космической академии им. А.Ф. Можайского. 2015. № 649. С. 16–22.
18. Кочегаров В.А., Фролов Г.А. Проектирование систем распределения информации. Марковские и немарковские модели. М.: Радио и связь, 1991. 216 с.
19. Буранова М.А., Карташевский В.Г. Анализ времени ожидания для узла сети типа G/D/1 при неточном знании параметров трафика // Информационные технологии и телекоммуникации. 2017. Т. 5. № 1. С. 24–33.
20. Букашкин С.А., Карташевский В.Г., Сапрыкин А.В. Анализ функционирования сетевого узла при неточном знании параметров трафика // DSPA: Вопросы применения цифровой обработки сигналов. 2017. Т. 7. № 2. С. 14–17.
21. Рыжиков Ю.И. Имитационное моделирование. Авторская имитация систем и сетей с очередями: Учебное пособие. СПб.: Издательство «Лань», 2019. 112 с.
22. Stefanov S.K. On the Basic Concepts of the Direct Simulation Monte Carlo Method. Physics of Fluids. 2019. vol. 31. no. 6. DOI: 10.1063/1.5099042.
23. Грешилов А.А. Анализ и синтез стохастических систем. Параметрические модели и конфлюентный анализ. М.: Радио и связь, 1990. 320 с.

24. Кузнецов В.П. Интервальные статистические модели. М.: Радио и связь, 1991. 452 с.
25. Левин В.И. Вычисления в условиях неопределенности с помощью интервальной математики // Донецкие чтения 2019: образование, наука, инновации, культура и вызовы современности. Материалы IV Международной научной конференции. Том 1 - Физико-математические и технические науки, Часть 2. 2019. С. 184–185.
26. Лазарев В.Л., Уваров Р.А. Организация адаптивного управления на основе информационных критериев. Мягкие измерения и вычисления. 2023. Т. 64. № 3. С. 46–57.
27. Zhao Y., Wu H., Yang C., Liu Z., Cheng Q. New Reliability Modeling Methods for Structural Systems with Hybrid Uncertainty. Quality and Reliability Engineering International. 2020. vol. 36. no. 6. pp. 1855–1871.
28. Гельфанд И.М., Шилов Г.Е. Обобщенные функции. Вып. 2. Пространства основных и обобщенных функций. М.: Физматгиз, 1958. 307 с.
29. Гончаренко В.А. Композиционный метод формирования аппроксимационных распределений с произвольной фазовой функцией // Труды СПИИРАН. 2016. Т. 3(46). С. 212–225.
30. He Q.-M., Liu B., Wu H. Continuous Approximations of Discrete Phase-Type Distributions and Their Applications to Reliability Models. Performance Evaluation. 2022. vol. 154. DOI: 10.1016/j.peva.2022.102284.
31. Sarada Y., Shenbagam R. Approximations of Availability Function Using Phase Type Distribution. Opsearch. 2022. vol. 59. pp. 1337–1351.
32. Wang L., Li Y., Qian Y., Luo X. A Parameter Estimation Method of Shock Model Constructed with Phase-Type Distribution on the Condition of Interval Data. Mathematical Problems in Engineering. 2020. vol. 2020. no. 11. DOI: 10.1155/2020/1424105.
33. Смагин В.А., Филимонихин Г.В. О моделировании случайных процессов на основе гипердельтного распределения // Автоматика и вычислительная техника. 1990. № 1. С. 25–31.
34. Буранова М.А., Карташевский В.Г. Рекурсивный подбор параметров гиперэкспоненциальных распределений при аппроксимации распределений с «тяжелыми хвостами». Труды учебных заведений связи. 2023. Т. 9. № 2. С. 40–46.
35. Докучаева А.Н. Статистическое имитационное моделирование систем массового обслуживания с параметрической неопределенностью посредством динамических упрощенных моделей // Системы управления и информационные технологии. 2019. № 1(75). С. 11–16.
36. Тарасов В.Н., Бахарева Н.Ф. Имитационное моделирование систем массового обслуживания на основе составных распределений – вероятностных смесей // T-Comm: Телекоммуникации и транспорт. 2023. Т. 17. № 3. С. 14–19.
37. Gofman M.V., Kornienko A.A., Glukharev M.L. A Method for Watermark Detection in Digital Audio Signals by Authorized Users // Automatic Control and Computer Sciences. 2021. vol. 55. no. 8. pp. 1005–1019.
38. Privalov A., Lukicheva V., Kotenko I., Saenko I. Increasing the sensitivity of the method of early detection of cyber-attacks in telecommunication networks based on traffic analysis by extreme filtering // Energies. 2020. vol. 13. no. 11.
39. Glukharev M., Solomatova M. Access Differentiation in Object-Oriented Databases Based on the Extended Object-Oriented Harrison–Ruzzo–Ullman Model // Automatic Control and Computer Sciences. 2020. vol. 54. pp. 1007–1012.

Гончаренко Владимир Анатольевич — канд. техн. наук, доцент кафедры, кафедра «информационные и вычислительные системы», Петербургский государственный университет путей сообщения Императора Александра I; преподаватель кафедры, кафедра информационно-вычислительных систем и сетей, ВКА им. А.Ф. Можайского. Область научных интересов: теория очередей, математическое моделирование, системы массового обслуживания с неопределенностью, информационная безопасность, устойчивость компьютерных сетей. Число научных публикаций — 111. vlango@mail.ru; улица Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)347-9524.

Хомоненко Анатолий Дмитриевич — д-р техн. наук, профессор кафедры, кафедра «информационные и вычислительные системы», Петербургский государственный университет путей сообщения Императора Александра I; профессор кафедры, кафедра математического и программного обеспечения, ВКА им. А.Ф. Можайского. Область научных интересов: численная теория массового обслуживания, моделирование сложных систем, программирование, операционные и информационные системы, системы массового обслуживания, надежность программного обеспечения, информационные системы, безопасность информационных систем. Число научных публикаций — 287. khomon@mail.ru; Московский проспект, 9, 190031, Санкт-Петербург, Россия; р.т.: +7(812)457-8356.

Абу Хасан Рахеб — аспирант, кафедры «информационные и вычислительные системы», Петербургский государственный университет путей сообщения Императора Александра I. Область научных интересов: информационные системы, хранилища данных, модели надежности. Число научных публикаций — 5. ragheb1997@yandex.ru; Московский проспект, 9, 190031, Санкт-Петербург, Россия; р.т.: +7(812)457-8356.

DOI 10.15622/ia.23.6.1

V. GONCHARENKO, A. KHOMONENKO, R. ABU KHASAN

A COMPOSITIONAL APPROACH TO THE SIMULATION OF QUEUING SYSTEMS WITH RANDOM PARAMETERS

Goncharenko V., Khomonenko A., Abu Khasan R. **A Compositional Approach to the Simulation of Queuing Systems with Random Parameters.**

Abstract. A general approach to modeling random service processes under conditions of disturbances and uncertainty of the initial data is substantiated. A compositional approach to constructing simulation models of queuing with parametric uncertainty based on phase-type distributions and phase functions is proposed. The calculation and comparison of the characteristics of the developed simulation models with analytical solutions were carried out to confirm their effectiveness and accuracy. The problems of uncertainty of the initial data and their impact on the modeling of service systems are highlighted. The importance of taking into account parametric uncertainty in simulation models is emphasized in order to increase their adequacy and applicability in practice. The study includes a description of a general approach to modeling random service processes with uncertainty, as well as methodological foundations for the application of phase distributions and functions in compositional modeling. Four classes of service models are considered, differing in the type of integral core and phase function, which makes it possible to implement a variety of random service processes, taking into account their characteristics and conditions of their occurrence. The analysis of a model with an exponential integral core and various types of phase functions is carried out, which demonstrates the flexibility and wide possibilities of the proposed compositional approach to the study and modeling of service systems. The results of simulation modeling are presented, confirming analytical studies and showing the applicability and effectiveness of the developed approach in the construction and analysis of models of service systems with random parameters. The practical significance of the compositional method for the design and modernization of information and computing systems at various stages of their development, taking into account the uncertainty of the initial data, is noted. The work is focused on the development of simulation methods for queuing systems and opens up new prospects for their research and optimization in conditions of uncertainty of initial parameters.

Keywords: compositional approach, integral kernel, simulation modeling, random parameter, parametric uncertainty, hyper-delta probability distribution, generalized function, phase-type distribution, uniformly exponential distribution, approximation, phase function, queueing systems.

References

1. Khomonenko A., Khalil M., Kassymova D. Probabilistic Models for Evaluating the Performance of Cloud Computing Systems with Web Interface. SPIIRAS Proceedings. 2016. vol. 6. no. 49. pp. 49–65.
2. Dudina O., Dudin A. Optimization of queueing model with server heating and cooling. Mathematics. 2019. vol. 7. no. 9. DOI: org/10.3390/math7090768.
3. Klimenok V., Dudin A., Dudina O., Kochetkova I. Queuing system with two types of customers and dynamic change of a priority. Mathematics. 2020. vol. 8(5). DOI: org/10.3390/math8050824.
4. Khomonenko A. Performance analysis of the multiprocessor systems with priority processing of heterogeneous requests flow. Avtomatika i Vychislitel'naya Tekhnika. 1991. no. 4. pp. 55–64.

5. Krasnov S., Lokhvitskiy V., Khabarov R. [Numerical analysis of multi-channel queuing systems with absolute priority based on phase approximation of the period of continuous employment]. Trudy Voyenno-kosmicheskoy akademii imeni A.F. Mozhayskogo – Proceedings of the Military Space Academy named after A.F. Mozhaisky. 2022. no. 682. pp. 7–20. (In Russ.).
6. Nazarov A., Strik J., Kvach A. A survey of recent results in finite-source retrial queues with collisions. Information technologies and mathematical modelling. Queueing Theory and Applications: 17th International Conference and 12th Workshop on Retrial Queues and Related Topics. 2018. vol. 912. pp. 1–15. DOI: 10.1007/978-3-319-97595-5_1.
7. Gaitonde J., Tardos E. Stability and learning in strategic queuing systems. Proceedings of the 21st ACM Conference on Economics and Computation. 2020. pp. 319–347.
8. Zayats O., Korenevskaya M., Ilyashenko A., Muliukha V. Prioritized Retrial Queueing Systems with Randomized Push-Out Mechanism. Informatics and Automation. 2024. vol. 23. no. 2. pp. 325–351. (In Russ.).
9. Buslenko N.P. Modelirovanie slozhnyh sistem [Modeling of complex systems]. M.: Nauka, 1978. 400 p. (In Russ.).
10. Ivnitskiy V. [On evaluating the accuracy of the simulation results of complex systems with inaccurate input information in the independent test scheme]. Izvestiya AN SSSR. Tekhnicheskaya kibernetika – Izvestia of the USSR Academy of Sciences. Technical cybernetics. 1974. no. 4. pp. 208–217. (In Russ.).
11. Law A. Simulation Modeling and Analysis, 6th Edition. McGraw Hill. 2024. 688 p.
12. Goncharenko V. [The formal apparatus of representation of stochastic processes of service with the disturbance and uncertainty parameters]. Trudy VKA imeni A.F. Mozhajskogo – Proceedings of Military Space Academy named A.F. Mozhaysky. 2015. vol. 648. pp. 13–18. (In Russ.).
13. Corlu C., Akcay A., Xie W. Stochastic simulation under input uncertainty: A review. Operations Research Perspectives. 2020. vol. 7. DOI: 10.1016/j.orp.2020.100162.
14. Shone R., Glazebrook K., Zografos K. Applications of stochastic modeling in air traffic management: Methods, challenges and opportunities for solving air traffic problems under uncertainty. European Journal of Operational Research. 2021. vol. 292. no. 1. pp. 1–26.
15. Xie W., Li C., Wu Y., Zhang P. A nonparametric Bayesian framework for uncertainty quantification in stochastic simulation. SIAM/ASA Journal on Uncertainty Quantification. 2021. vol. 9. no. 4. pp. 1527–1552.
16. Goncharenko V. [Models and methods of queueing systems analysis with parametric uncertainty]. Intellektual'nye tekhnologii na transporte – Intellectual Technologies on Transport. 2017. no. 4. pp. 5–11. (In Russ.).
17. Goncharenko V. [Modelling and estimation of characteristics of random flows of events in computer networks under parametric uncertainty]. Trudy VKA imeni A.F. Mozhajskogo – Proceedings of Military Space Academy named A.F. Mozhaysky. 2015. vol. 649. pp. 16–22. (In Russ.).
18. Kochegarov V., Frolov G. Proektirovanie sistem raspredelenija informacii. Markovskie i nemarkovskie modeli. [Designing of systems of information distribution.Markovian and non-Markovian models]. M.: Radio i svjaz', 1991. 216 p. (In Russ.).
19. Buranova M., Kartashevskiy V. [The Analysis of the Latency Period for Knot of Network of the G/D/1 Type at Inaccurate Knowledge of Parameters of the Traffic]. Informatsionnye tekhnologii i telekommunikatsii – Inf. Technol. Telecommunications. 2017. vol. 5. no. 1. pp. 24–33. (In Russ.).

20. Bukashkin S., Kartashevskij V., Saprykin A. [Analysis of the functioning of a network node with inaccurate knowledge of traffic parameters]. DSPA: Voprosy primeneniya cifrovoj obrabotki signalov – DSPA: Issues of application of digital signal processing. 2017. vol. 7. no. 2. pp. 14–17. (In Russ.).
21. Ryzhikov Yu. Imitacionnoe modelirovanie. Avtorskaya imitaciya sistem i setej s ocheredyami. [Simulation modeling. Author's simulation of systems and networks with queues]. St. Petersburg: Lan Publishing House, 2019. 112 p. (In Russ.).
22. Stefanov S. On the Basic Concepts of the Direct Simulation Monte Carlo Method. Physics of Fluids. 2019. vol. 31. no. 6. DOI: 10.1063/1.5099042.
23. Greshilov A. Analiz i sintez stohasticheskih sistem. Parametricheskie modeli i konflyuentnyj analiz [Analysis and synthesis of stochastic systems. Parametric models and confluent analysis]. M.: Radio i svyaz', 1990. 320 p. (In Russ.).
24. Kuznetsov V. Interval'nye statisticheskie modeli [Interval statistical models]. M.: Radio i svyaz', 1991. 452 p. (In Russ.).
25. Levin V. [Calculations under uncertainty using interval mathematics] Doneckie chteniya 2019: obrazovanie, nauka, innovacii, kul'tura i vyzovy sovremennosti. Materialy IV Mezhdunarodnoj nauchnoj konferencii [Donetsk Readings 2019: education, science, innovation, culture and modern challenges. Materials of the IV International Scientific Conference]. 2019. pp. 184–185. (In Russ.).
26. Lazarev V., Uvarov R. [The organization of adaptive management based on information criteria]. Myagkie izmereniya i vychisleniya – Soft measurements and calculations. 2023. vol. 64. no. 3. pp. 46–57. (In Russ.).
27. Zhao Y., Wu H., Yang C., Liu Z., Cheng Q. New Reliability Modeling Methods for Structural Systems with Hybrid Uncertainty. Quality and Reliability Engineering International. 2020. vol. 36. no. 6. pp. 1855–1871.
28. Gel'fand M., Shilov G. Obobshhennye funkcii. Vyp. 2. Prostranstva osnovnyh i obobshhennyh funkcij [Generalized functions. Vol 2. Spaces of the test and generalized functions.]. M.: Fizmatgiz. 1958. 307 p. (In Russ.).
29. Goncharenko V. [Composite Method of Forming Approximating Distributions with an Arbitrary Phase Function]. Trudy SPIIRAN – SPIIRAS Proc. 2016. vol. 46. no. 3. pp. 212–225. (In Russ.).
30. He Q.-M., Liu B., Wu H. Continuous Approximations of Discrete Phase-Type Distributions and Their Applications to Reliability Models. Performance Evaluation. 2022. vol. 154. DOI: 10.1016/j.peva.2022.102284.
31. Sarada Y., Shenbagam R. Approximations of Availability Function Using Phase Type Distribution. Opsearch. 2022. vol. 59. pp. 1337–1351.
32. Wang L., Li Y., Qian Y., Luo X. A Parameter Estimation Method of Shock Model Constructed with Phase-Type Distribution on the Condition of Interval Data. Mathematical Problems in Engineering. 2020. vol. 2020. no. 11. DOI: 10.1155/2020/1424105.
33. Smagin V., Filimonihin G. [About modelling of stochastic processes on a basis of hyperdelta distribution]. Avtomatika i vychislitel'naja tehnika – Automation and Computer Engineering. 1990. no. 1. pp. 25–31. (In Russ.).
34. Buranova M., Kartashevsky V. [Recursive selection of parameters of hyperexponential distributions in the approximation of distributions with "heavy tails"]. Trudy uchebnyh zavedenij svyazi – Proceedings of educational institutions of communication. 2023. vol. 9. no. 2. pp. 40–46. (In Russ.).
35. Dokuchaeva A. [Monte-Carlo Simulation of the Queuing Systems with Parametric Uncertainty by the Dynamic Simplified Models]. Sistemy upravleniya i informacionnye tekhnologii – Control systems and information technologies. 2019. no. 1(75). pp. 11–16. (In Russ.).

36. Tarasov V., Bakhareva N. [Simulation modeling of queuing systems based on composite distributions – probabilistic mixtures]. T-Comm: Telecommunications and Transport – T-Comm: Telekommunikacii i transport. 2023. vol. 17. no. 3. pp. 14–19.
37. Gofman M., Kornienko A., Glukharev M. A Method for Watermark Detection in Digital Audio Signals by Authorized Users. Automatic Control and Computer Sciences. 2021. vol. 55. no. 8. pp. 1005–1019.
38. Privalov A., Lukicheva V., Kotenko I., Saenko I. Increasing the sensitivity of the method of early detection of cyber-attacks in telecommunication networks based on traffic analysis by extreme filtering. Energies. 2020. vol. 13. no. 11.
39. Glukharev M., Solomatova M. Access Differentiation in Object-Oriented Databases Based on the Extended Object-Oriented Harrison–Ruzzo–Ullman Model. Automatic Control and Computer Sciences. 2020. vol. 54. pp. 1007–1012.

Goncharenko Vladimir — Ph.D., Associate professor of the department, Department of information and computing systems, Emperor Alexander I St. Petersburg State Transport University; Lecturer of the department, Information and computing systems and networks department, Mozhaisky Military Space Academy. Research interests: queue theory, mathematical modeling, queueing systems with uncertainty, information security, stability of computer networks. The number of publications — 111. vlango@mail.ru; 13, Zhdanovskaya St., 197198, St. Petersburg, Russia; office phone: +7(812)347-9524.

Khomonenko Anatoly — Ph.D., Dr.Sci., Professor of the department, Department of information and computing systems, Emperor Alexander I St. Petersburg State Transport University; Professor of the department, Department of mathematical and software engineering, Mozhaisky Military Space Academy. Research interests: numerical theory of queuing, modeling of complex systems, programming, operating and information systems, queuing systems, software reliability, information systems, information system security. The number of publications — 287. khomon@mail.ru; 9, Moskovsky Ave., 190031, St. Petersburg, Russia; office phone: +7(812)457-8356.

Abu Khasan Rakheb — Graduate student, Department of information and computing systems, Emperor Alexander I St. Petersburg State Transport University. Research interests: information systems, data warehouses, reliability models. The number of publications — 5. ragheb1997@yandex.ru; 9, Moskovsky Ave., 190031, St. Petersburg, Russia; office phone: +7(812)457-8356.

УДК 519.872.5 **DOI** 10.15622/ia.23.6.2

С.З. Куракин, А.Ю. Онуфрей, А.В. Разумов

ИССЛЕДОВАНИЕ ВАРИАНТОВ ПОСТРОЕНИЯ ИНФОРМАЦИОННО-УПРАВЛЯЮЩИХ СИСТЕМ НА ОСНОВЕ СЕТЕВЫХ МОДЕЛЕЙ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

Куракин С.З., Онуфрей А.Ю., Разумов А.В. **Исследование вариантов построения информационно-управляющих систем на основе сетевых моделей систем массового обслуживания.**

Аннотация. Одним из направлений дальнейшего совершенствования и повышения эффективности применения технических объектов при решении ими целевых задач является применение информационно-управляющих систем (ИУС) для управления сложными техническими объектами. Существующие современные ИУС представляют собой комплекс аппаратно-программных средств, предназначенных для сбора, обработки и хранения информации и управления. В условиях наличия большого количества информации, противоречивых факторов, влияющих на качество управления, принятие обоснованных и своевременных решений в процессе управления невозможно без применения ИУС. Разрабатываемые ИУС, как правило, являются специализированными системами и проектируются для решения конкретных задач. В связи с этим разработка и проектирование ИУС должны проводиться с учетом взаимосвязи с целевыми показателями и особенностями объектов управления, а также результатами всестороннего анализа информации о параметрах ИУС, влияющих на показатели их эффективности. Использование математических моделей для исследования вариантов построения ИУС является основой проектирования и разработки устройств и подсистем ИУС. Разрабатываемые в настоящее время модели ИУС, как правило, позволяют проводить исследования для одностадийных процессов управления с наличием в системе однотипных объектов обслуживания. В то же время современные технические объекты и системы управления представляют собой сложные комплексы с циклически повторяющимися процессами управления разнотипными средствами. Как правило, в таких комплексах имеется набор параллельно работающих устройств (каналов управления), обеспечивающих управление разнотипных объектов на различных стадиях обработки информации. В этом случае структуру ИУС необходимо представлять в виде многофазной многоканальной технической системы, в которой происходит процесс одновременного управления несколькими объектами различных типов. В связи с этим целью статьи является разработка и исследование математической модели ИУС с двумя фазами обработки и наличием определенного количества обслуживающих разнотипных устройств. Основой модели является многофазная сетевая модель системы массового обслуживания. Исследование модели позволяет выбрать вариант построения ИУС, в частности выбрать оптимальное количество каналов обработки для различных типов объектов по критерию оптимальности с учетом ограничений по стоимости и времени обслуживания. Разработан алгоритм выбора варианта построения ИУС и приведен пример расчета количества каналов обработки в двухфазной системе при управлении тремя типами объектов.

Ключевые слова: информационно-управляющая система, математическое моделирование, сетевая многоканальная система массового обслуживания, выбор варианта построения, критерии оптимальности.

1. Введение. В настоящее время информационно-управляющие системы (ИУС) находят все более широкое применение в системах управления сложными техническими объектами. Современные ИУС представляют собой комплекс аппаратно-программных средств, предназначенных для сбора, обработки и хранения информации и управления. Основное предназначение ИУС заключается в обеспечении требуемой эффективности работы объектов управления при решении ими целевых задач. В условиях наличия большого количества информации, противоречивых факторов, влияющих на качество управления, принятие обоснованных и своевременных решений в процессе управления невозможно без применения ИУС. Разрабатываемые в настоящее время ИУС являются в большинстве своём специализированными системами и проектируются для решения конкретных задач независимо от того, что будет лежать в основе их реализации: специализированный микроконтроллер, вычислительная сеть или иные вычислительные средства. В связи с этим разработка и проектирование ИУС является важной научно-технической задачей, требующей всестороннего анализа информации об объектах управления, взаимосвязи элементов ИУС в процессе функционирования, структурных параметров ИУС и внешних факторов, влияющих на показатели эффективности управления.

Данному вопросу посвящено довольно большое количество работ отечественных и зарубежных авторов. Так в работах [1 – 5] рассматриваются вопросы построения ИУС и методического обеспечения для организации проектирования. В работах [1, 2] предложен научно-методический аппарат обоснования структуры информационно-управляющих систем, проведена вербальная и формализованная постановки задачи исследования, определены временные и стоимостные ограничения при выборе решений по созданию ИУС, предложены алгоритмы решения задач, основанные на применении методов оптимизации. В работах [3 – 5] показаны особенности проектирования ИУС на примере комплексов управления БПЛА [3], где предложен метод синтеза информационно-управляющей системы на основе сетецентрического подхода и представлена структура информационно-управляющей системы управления лесохозяйственным комплексом с использованием подсистемы поддержки принятия решения [4], рассматривается подходы к проведению синтеза распределенных ИУС [5]. В работах [6 – 8] рассматриваются вопросы, связанные с непосредственной разработкой ИУС. Предлагается учитывать при построении ИУС надежность оборудования систем управления и использовать динамическую

избыточности ресурсов в процессе функционирования [6, 7]. В работе [8] для оценки функциональной устойчивости элементов структуры ИУС предлагается использовать теорию полумарковских процессов. Часть работ посвящена анализу и оценке эффективности применения ИУС [9 – 11]. Предлагается рассматривать ИУС, как сложную техническую систему, при оценке эффективности которой необходимо учитывать иерархию построения и показатели эффективности процессов управления на различных уровнях иерархии ИУС [9, 10].

В [11] предложено описывать СУ в виде ориентированного графа и способ формализации задачи выбора варианта построения ИУС. Большое количество работ посвящено моделированию процессов управления в ИУС и разработки моделей расчета показателей результативности применения ИУС [12 – 32]. В работах [12 – 14] разработаны математические модели многостадийных процессов обслуживания произвольного количества заданий в конвейерных системах, а в работе [13] предложена сетевая модель процесса управления образовательной деятельностью вуза.

В работах [15 – 17] предлагается использовать для моделирования математический аппарат СМО на примере одноканальной системы массового обслуживания с приоритетами и повторным обслуживанием заявок [15]. Такие системы моделируют многие реальные ситуации, в частности, телекоммуникационные системы с протоколами множественного доступа при наличии коллизий CSMA/CD [16]. Проведены исследования различных способов формирования порядков выполнения пакетов заданий в многостадийных системах [17].

В работах [18 – 24] исследуются вопросы выбора пропускных способностей каналов связи. Разработана аналитическая модель транспортной сети связи, которая предполагает разбиение исходной сети на отдельные фрагменты, анализируемые независимо друг от друга с помощью систем массового обслуживания [18]. Проведенный анализ аналитических моделей сетей связи, основанных на математическом аппарате сетей массового обслуживания [19 – 24], показал, что они не позволяют учитывать особенности переноса трафика различного приоритета. В то же время метод декомпозиции, представленный в [19], позволяет описать исходную сеть связи с коммутацией пакетов в виде множества взаимодействующих систем массового обслуживания (СМО), входные и выходные процессы которых связаны между собой на уровне параметров.

В работах [25 – 30] предлагаются методы совместной обработки разнородных исходных данных в ЛВС. Описываются модели и методы построения сети в условиях недостатка данных, дается описание структуры ЛВС, включающее сведения об управляемых элементах и связях иерархии и передачи данных между ними. Работы [31, 11] посвящены выбору вариантов построения технических систем.

В [31] предлагается осуществлять выбор структуры на основе кластеризации Парето оптимальных решений в многомерном критериальном пространстве с использованием генетических алгоритмов, а в [11] выбор структуры проводится на основе двухступенчатой процедуры оптимизации, включающей выбор оптимального базового варианта на основе аналитического моделирования и дальнейшего его усовершенствования на имитационных моделях в соответствии с выбранным критерием оптимальности.

Анализ представленных работ показывает, что для исследования вариантов построения ИУС возможны различные подходы и методы моделирования. Большинство из них соответствуют цели исследования и особенностям исследуемых объектов. В данной статье для моделирования процессов управления в ИУС с двумя этапами обработки данных предлагается использовать двухфазную сетевую модель системы массового обслуживания (СМО). Обусловлено это особенностями обслуживания объектов в перспективных ИУС, в которых порядок обслуживания заявок представляет собой несколько последовательно повторяющихся циклов обслуживания с выполнением однотипных задач. Поэтому для моделирования таких процессов наиболее предпочтительным математическим аппаратом исследования являются сетевые модели СМО.

В связи с этим целью статьи является обоснование вариантов построения ИУС на основе моделирования процессов её функционирования в виде двухфазной сетевой модели СМО.

Для решения поставленной задачи необходимо провести формализованную постановку задачи исследования.

2. Постановка задачи выбора варианта построения информационно-управляющей системы. Рассмотрим процесс проектирования информационно-управляющей системы *F,* предназначенной для обслуживания объектов в области *D* за заданное время $T_{зад}$. Объекты в зависимости от своих характеристик могут относиться к определенному *i*-му типу ($i = \overline{1,I}$) и с различными интервалами времени входят в область *D.* При этом за время $T_{зад}$

в область D поступит множество объектов, количество которых характеризуется вектором $M = \{m_i\}, i = \overline{1,\ I}$, где m_i – математическое ожидание количества объектов i-го типа, I – количество разных типов объектов.

Обслуживание объектов производится в два этапа. На первом этапе объекты обслуживаются в подсистеме S, в которой решается задача обнаружения и распознавания объектов, а на втором этапе – в подсистеме Z, в которой решается задача обработки данных об объектах. При этом выходной поток обслуженных объектов подсистемы S является входным потоком для подсистемы Z.

Подсистемы S и Z состоят из определенного количества обслуживающих средств. При этом в подсистеме S все обслуживающие средства подразделяются на группы однотипных средств и характеризуются вектором $K^S = \{k_j^S\},\ j = \overline{1,\ J}$, где k_j^S – количество однотипных средств j-го типа, отличающихся по своим характеристикам и показателям качества обслуживания, J – количество типов обслуживающих средств в подсистеме S. В подсистеме Z все обслуживающие средства также подразделяются на группы однотипных средств и характеризуются вектором $K^Z = \{k_n^Z\}$, $n = \overline{1,\ N}$, где k_n^Z – количество однотипных средств n-го типа, отличающихся по своим характеристикам и показателям качества обслуживания, N – количество типов обслуживающих средств в подсистеме Z.

Заданы допустимые затраты $C_{\text{доп}}^S$ на создание подсистемы S и допустимые затраты $C_{\text{доп}}^Z$ на создание подсистемы Z.

Требуется выбрать вариант состава и количества обслуживающих устройств в подсистемах S и Z, при которых обеспечивается обслуживание максимального количества M объектов в области D за заданное время $T_{\text{зад}}$, что, в свою очередь, может быть обеспечено максимизацией пропускной способности системы.

Данная задача представляет собой оптимизационную задачу поиска варианта построения системы F, в которой в качестве критерия выбора оптимального варианта целесообразно использовать показатель $W(K^S, K^Z)$ пропускной способности системы, обеспечивающий максимальное количество обслуженных объектов в области D. Ограничениями при решении этой задачи являются затраты на разработку системы, включающие допустимые затраты $C_{\text{доп}}^S$ на разработку подсистемы S и допустимые затраты $C_{\text{доп}}^Z$ на разработку подсистемы Z, а также допустимое время $T_{\text{зад}}^S$ на обслуживание в подсистеме S и допустимое время $T_{\text{зад}}^Z$ на обслуживание в подсистеме Z.

С учетом ограничений задача выбора в формализованном виде может быть представлена следующим образом:

$$F(S,Z)\colon v^* = arg \max_{v \in V} W(K_v^S, K_v^Z) \quad (1)$$

при ограничениях

$$C_v^S(K_v^S) \le C_{\text{доп}}^S ;\ \ C_v^Z(K_v^Z) \le C_{\text{доп}}^Z ; \quad (2)$$

$$T_{\text{обсл}}^S(K_v^S) \le T_{\text{зад}}^S ;\ \ T_{\text{обсл}}^Z(K_v^Z) \le T_{\text{зад}}^Z , \quad (3)$$

где v – номер варианта построения подсистем S и Z ($v = \overline{1,V}$);

V – общее количество вариантов построения подсистем S и Z;

K_v^S – значение вектора K^S для v-го варианта построения ИУС;

K_v^Z – значение вектора K^Z для v-го варианта построения ИУС.

Для решения этой задачи целесообразно провести декомпозицию общей задачи на две подзадачи: подзадачу выбора варианта построения подсистемы S и подзадачу выбора варианта построения подсистемы Z.

2.1. Формализация задачи выбора состава обслуживающих средств подсистемы *S*. На первом этапе обслуживание объектов в информационно-управляющей системе F производится в подсистеме S.

Подсистема S состоит из множества обслуживающих средств, характеризуемых вектором K^S, элементами которого являются k_j^S – количества однотипных средств j-го типа ($j = \overline{1, J}$), отличающихся по своим характеристикам, заданных вектором характеристик $R = \{R_j\}$.

Заданы стоимости обслуживающих средств по каждому j-му типу средств C_j^S и допустимые затраты $C_{\text{доп}}^S$ на проектирование подсистемы S.

Входным потоком объектов на обслуживание для подсистемы S является входной поток объектов i-го типа, поступающих из области D. Каждый объект i-го типа описывается вектором характеристик $B = \{B_i\}$ и может быть обслужен средством j-го типа подсистемы S с вероятностью $P_{ij}^S,\ i = \overline{1,I},\ j = \overline{1,J}$.

Требуется определить вариант $v^* \in V$ построения подсистемы S, характеризуемый вектором K_{v*}^S, который определяет состав и количество средств каждого типа $k_j^S,\ j = \overline{1, J}$ и обеспечивает

максимальную пропускную способность $W(K_v^S, R, B)$ подсистемы S при обслуживании объектов из области D:

$$F(S^*): v^* = arg \max_{v' \in V} W(K_v^S, R, B) \tag{4}$$

при ограничениях

$$C^S = \sum_{j=1}^{J} C_j^S k_j^S \le C_{\text{доп}}^S ; \tag{5}$$

$$T_{i\,\text{обсл}}^S(K_v^S, R, B) \le T_{\text{зад}}^S, \qquad \forall\, i = \overline{1,I}. \tag{6}$$

Рассмотрим формализацию задачи выбора состава обслуживающих средств для подсистемы Z.

2.2. Формализация задачи выбора состава обслуживающих средств подсистемы *Z*. На втором этапе обслуживание объектов в информационно-управляющей системе F производится в подсистеме Z.

Подсистема Z состоит из множества обслуживающих средств, характеризуемых вектором K^Z, элементами которого являются k_n^Z – количества однотипных средств n-го типа ($n = \overline{1,\ N}$), отличающихся по своим характеристикам, заданных вектором характеристик $H = \{H_n\}$.

Заданы стоимости обслуживающих средств по каждому n-му типу средств C_n^Z и допустимые затраты $C_{\text{доп}}^Z$ на проектирование подсистемы Z.

Входным потоком объектов на обслуживание для подсистемы Z является выходной поток обслуженных объектов из подсистемы S. Каждый объект i-го типа описывается вектором характеристик $B = \{B_i\}$ и может быть обслужен средством n-го типа подсистемы Z с вероятностью $P_{in}^Z,\ i = \overline{1,I},\ n = \overline{1,N}$.

Требуется определить вариант $v^{**} \in V$ построения подсистемы Z, характеризуемый вектором K_v^Z, который определяет состав и количество средств каждого типа $k_n^Z,\ n = \overline{1,\ N}$ и обеспечивает максимальную пропускную способность $W(K_v^Z, H, B)$ подсистемы Z при обслуживании объектов, поступивших после обслуживания в подсистеме S:

$$F(Z^*): v^{**} = arg \max_{v'' \in V} W(K_v^Z, H, B) \quad (7)$$

при ограничениях

$$C^Z = \sum_{n=1}^{N} C_n^Z k_n^Z \le C_{\text{доп}}^Z ; \quad (8)$$

$$T_{i\,\text{обсл}}^Z(K_v^Z, H, B) \le T_{\text{зад}}^Z, \ \forall\, i = \overline{1, I}. \quad (9)$$

Решение представленных формализованных постановок задач выбора состава обслуживающих средств в системе F представляет собой сложную комбинаторную задачу. Поэтому целесообразно рассмотреть подходы к решению этой задач и на основе моделирования процессов обслуживания объектов в ИУС с помощью аппарата теории массового обслуживания.

3. Моделирование процессов обслуживания объектов в ИУС. Рассмотрим процессы обслуживания объектов в системе F в терминах СМО с учетом ранее представленной формализованной постановки задачи. Учитывая, что процесс обслуживания представляет собой двухэтапную процедуру, то для её моделирования, целесообразно использовать сетевую двухфазную СМО, каждая фаза которой, в свою очередь, состоит из узлов определенного типа, что позволяет имитировать процесс обработки потока заявок в проектируемой ИУС (рисунок 1).

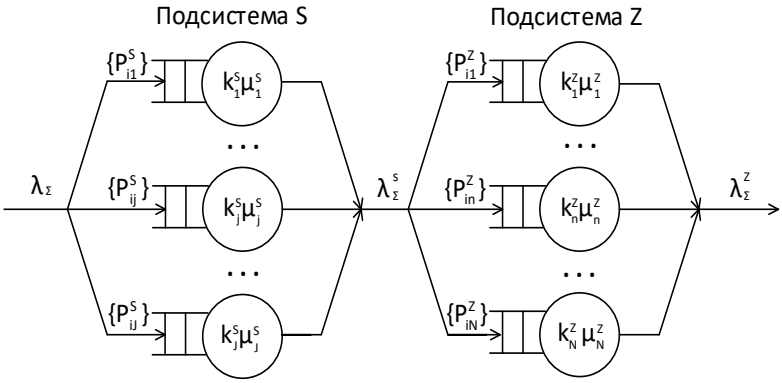
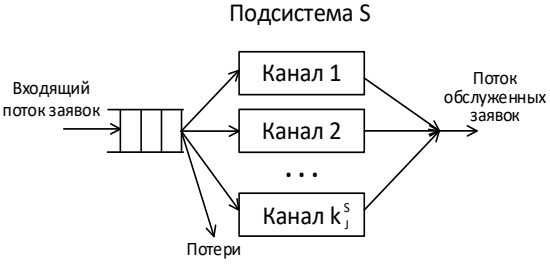
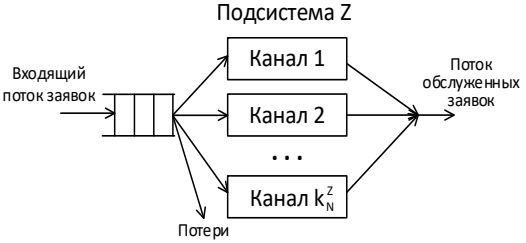


Рис. 1. Структура сетевой двухфазной многоканальной СМО, имитирующей процесс обработки потока заявок в проектируемой ИУС

В данной сетевой двухфазной СМО каждый узел представляет собой многоканальную СМО с очередью, в которой каналы моделируют процессы обслуживания однотипными устройствами в составе подсистем *S* и *Z* (рисунок 2).



а)



б)

Рис. 2. Структура узла в виде многоканальной СМО в составе сетевой СМО: а) для подсистемы *S*; б) для подсистемы *Z*

В соответствии с постановкой задачи на вход ИУС, представленной в виде сетевой двухфазной многоканальной СМО, поступает суммарный входной поток λ_Σ заявок, состоящий из заявок *i*-го типа:

$$\lambda_\Sigma = \lambda_1 + \lambda_2 + \cdots + \lambda_i + \cdots \lambda_I, \tag{10}$$

где λ_i – интенсивность входного потока заявок *i*-го типа ($i = \overline{1,\ I}$);

$p_i = \lambda_i/\lambda_\Sigma$ – вероятность прихода заявки *i*-го типа ($i = \overline{1,\ I}$);

$p_1 + p_2 + \cdots + p_I = 1$ – условие нормировки.

Первая фаза сетевой СМО имитирует процесс функционирования подсистемы *S*, при этом в её состав входят узлы *j*-го типа ($j = \overline{1,\ J}$), каждый из которых представляет собой

многоканальную СМО с очередью. Каждый канал обслуживания в многоканальной СМО j-го типа ($j = \overline{1,\ J}$) моделирует работу однотипных информационных средств подсистемы S. Количество каналов обслуживания СМО соответствует количеству информационных средств j-го типа ($j = \overline{1,\ J}$).

При этом приведенная интенсивность обслуживания заявок в многоканальной СМО узлом j-го типа ($j = \overline{1,\ J}$) зависит от типа обслуживаемой заявки и определяется по формуле:

$$\mu_j^S = \sum_{i=1}^{I} P_{ij}^S k_j^S \mu_{ij}^S , \tag{11}$$

где P_{ij}^S – вероятность прихода заявки i-го типа ($i = \overline{1,\ I}$) в узел j-го типа ($j = \overline{1,\ J}$) в подсистеме S;

k_j^S – количество каналов обслуживания узла j-го типа ($j = \overline{1,\ J}$) в подсистеме S;

μ_{ij}^S – интенсивность обслуживания заявок i-го типа одним каналом обслуживания узла j-го типа в подсистеме S.

Вторая фаза сетевой СМО имитирует процесс функционирования подсистемы Z, при этом в её состав входят узлы n-го типа ($n = \overline{1,\ N}$), каждый из которых представляет собой многоканальную СМО с очередью. Каждый канал обслуживания в узле n-го типа ($n = \overline{1,\ N}$) моделирует работу однотипных обслуживающих средств подсистемы Z. Количество каналов обслуживания в узле соответствует количеству обслуживающих средств n-го типа ($n = \overline{1,\ N}$).

При этом приведенная интенсивность обслуживания заявок в многоканальной СМО узлом n-го типа ($n = \overline{1,\ N}$) зависит от типа обслуживаемой заявки и определяется по формуле:

$$\mu_n^Z = \sum_{i=1}^{I} P_{in}^Z k_n^Z \mu_{in}^Z , \tag{12}$$

где P_{ij}^Z – вероятность прихода заявки i-го типа ($i = \overline{1,\ I}$) в узел n-го типа ($n = \overline{1,\ N}$) в подсистеме Z;

k_n^Z – количество каналов обслуживания узла n-го типа ($n = \overline{1,\ N}$) в подсистеме Z;

μ_{in}^{Z} – интенсивность обслуживания заявок *i*-го типа одним каналом обслуживания узла *n*-го типа в подсистеме *Z*.

Дисциплина обслуживания в узлах каждой из подсистем *S* и *Z* задается с учетом ограничений на время пребывания заявки в системе, при которых пришедшая на обслуживание заявка получает отказ при отсутствии наличия свободных каналов или, когда её время ожидания в очереди превышает допустимое значение.

Рассмотрим показатели эффективности многоканальных СМО, которые моделируют процессы обслуживания заявок в узлах сетевых СМО, моделирующих процессы функционирования подсистем *S* и *Z*.

В качестве критерия выбора количества каналов в узлах сетевых СМО целесообразно использовать показатель загрузки каналов обслуживания. При этом загрузка каналов обслуживания каждого из узлов *j*-го типа в подсистеме *S* определяется по формуле:

$$\rho_j^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_j^S \mu_{ij}^S}, \tag{13}$$

где μ_{ij}^S – интенсивность обслуживания заявок *i*-го типа одним каналом обслуживания узла *j*-го типа в подсистеме *S*.

Загрузка каналов обслуживания каждого из узлов *n*-го типа в подсистеме *Z* определяется по формуле:

$$\rho_n^Z = \sum_{i=1}^{I} \frac{\lambda_i}{k_n^Z \mu_{in}^Z}, \tag{14}$$

где μ_{in}^Z – интенсивность обслуживания заявок *i*-го типа одним каналом обслуживания узла *n*-го типа в подсистеме *Z*.

Вероятность отказа в обслуживании заявок зависит от распределения вероятностей нахождения многоканальных СМО в определенных состояниях системы, которые нумеруются по числу заявок, находящихся в очереди и на обслуживании в системе. Далее будем рассматривать предельные вероятности состояний системы для стационарного режима работы.

В соответствии с [32] предельные вероятности состояний системы определяются следующими выражениями.

Вероятность нахождения многоканальных СМО (узлов), входящих в состав сетевых СМО, в свободном состоянии определяется:

– для узлов j-го типа в подсистеме S:

$$P_0^{Sj} = 1 \Big/ \sum_{r=1}^{k_j^S} \left(1 + \frac{\rho_j^r}{r!}\right); \tag{15}$$

– для узлов n-го типа в подсистеме Z:

$$P_0^{Zn} = 1 \Big/ \sum_{r=1}^{k_n^Z} \left(1 + \frac{\rho_n^r}{r!}\right). \tag{16}$$

Вероятность отказа в обслуживании заявок в каждом из узлов сетевых СМО равна вероятности того, что все каналы соответствующего узла заняты, и определяется:

– для узлов j-го типа в подсистеме S:

$$P_{\text{отк}}^{Sj} = \frac{\rho_j^{k_j^S}}{k_j^S!} P_0^{Sj}; \tag{17}$$

– для узлов n-го типа в подсистеме Z:

$$P_{\text{отк}}^{Zn} = \frac{\rho_n^{k_n^Z}}{k_n^Z!} P_0^{Zn}. \tag{18}$$

После определения вероятности отказа в обслуживании заявок в узлах сетевых СМО можно определить относительную пропускную способность:

– для узлов j-го типа в подсистеме S:

$$Q^{Sj} = 1 - P_{\text{отк}}^{Sj}; \tag{19}$$

– для узлов n-го типа в подсистеме Z:

$$Q^{Zn} = 1 - P_{\text{отк}}^{Zn} . \quad (20)$$

Так как узлы, входящие в состав сетевых СМО и представляющие собой отдельные многоканальные СМО, функционируют параллельно, то оценку относительной пропускной способности подсистем S и Z можно получить, используя соотношения для схемы параллельного соединения:

– для подсистемы S:

$$P_{\text{отк}}^{S} = 1 \Big/ \sum_{j=1}^{J} \left(1/P_{\text{отк}}^{Sj}\right) ; \quad (21)$$

– для подсистемы Z:

$$P_{\text{отк}}^{Z} = 1 \Big/ \sum_{n=1}^{N} \left(1/P_{\text{отк}}^{Zn}\right) . \quad (22)$$

При этом интенсивность выходного потока определяется:

– из подсистемы S:

$$\lambda_{\Sigma}^{S} = \lambda_{\Sigma}(1 - P_{\text{отк}}^{S}); \quad (23)$$

– из подсистемы Z:

$$\lambda_{\Sigma}^{Z} = \lambda_{\Sigma}(1 - P_{\text{отк}}^{S})(1 - P_{\text{отк}}^{Z}). \quad (24)$$

Абсолютные пропускные способности узлов рассчитываются в соответствии с выражениями:

– для подсистемы S:

$$A^{S} = \lambda_{\Sigma}^{S} \cdot Q^{S}; \quad (25)$$

– для подсистемы Z:

$$A^Z = \lambda_\Sigma^Z \cdot Q^Z. \quad (26)$$

Значение показателя стоимости определяется:

– для подсистемы *S*:

$$C^S = \sum_{j=1}^{J} k_j^S c_j^S ; \quad (27)$$

– для подсистемы *Z*:

$$C^Z = \sum_{n=1}^{N} k_n^Z c_n^Z . \quad (28)$$

Таким образом, в соответствии с предложенным описанием системы F в терминах СМО поставленную задачу выбора вариантов построения ИУС с учетом ограничений в формализованном виде можно представить в следующем виде:

$$F(S,Z): v^* = arg \max_{v \in V} W(K_v^S, K_v^Z, \lambda_\Sigma, \mu^S, \mu^Z) \quad (29)$$

при ограничениях

$$T_{\text{обсл}}^S(K_v^S, \lambda_\Sigma, \mu^S, \mu^Z) \le T_{\text{зад}}^S ; \quad (30)$$

$$T_{\text{обсл}}^Z(K_v^Z, \lambda_\Sigma, \mu^S, \mu^Z) \le T_{\text{зад}}^Z ; \quad (31)$$

$$C^S = \sum_{j=1}^{J} k_j^S c_j^S \le C_{\text{доп}}^S ; \quad (32)$$

$$C^Z = \sum_{n=1}^{N} k_n^Z c_n^Z \le C_{\text{доп}}^Z ; \quad (33)$$

$$\rho_j^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_j^S \mu_{ij}^S} \le \rho_{\text{доп}}^S,\ \forall j = \overline{1,J}\ ; \tag{34}$$

$$\rho_n^Z = \sum_{i=1}^{I} \frac{\lambda_i}{k_n^Z \mu_{in}^Z} \le \rho_{\text{доп}}^Z,\ \forall n = \overline{1,N}\ , \tag{35}$$

где v – номер варианта построения подсистем S и Z ($v = \overline{1,\ V}$);

V – общее количество вариантов построения подсистем S и Z.

Целью исследования является определение вектора $K^S = \{k_j^S\}$, $j = \overline{1,\ J}$, элементы k_j^S которого определяют количество каналов обслуживания узлов j-го типа в подсистеме S и вектора $K^Z = \{k_n^Z\}$, $n = \overline{1,\ N}$, элементы k_n^Z которого определяют количество каналов обслуживания узлов n-го типа в подсистеме Z, при которых обеспечивается максимальная пропускная способность этих подсистем и обслуживание максимального количества объектов в области D за заданное время $T_{\text{зад}}$.

Данная задача предполагает некоторую процедуру поиска оптимального решения, которая может быть представлена в виде алгоритма выбора состава средств обслуживания в указанных подсистемах S и Z.

4. Алгоритм выбора состава обслуживающих средств в подсистемах *S* и *Z*. Алгоритм выбора состава обслуживающих средств в подсистемах S и Z представлен на рисунке 3.

Исходными данными для алгоритма являются:

– суммарный входной поток λ_Σ заявок, состоящий из заявок i-го типа и заданный вектором $\lambda_\Sigma = \{\lambda_i\}$, $i = \overline{1,\ I}$ интенсивностей поступления заявок i-го типа;

– матрица $\mu^S = \{\mu_{ij}^S\}$, $i = \overline{1,\ I}$, $j = \overline{1,\ J}$, состоящая из элементов μ_{ij}^S интенсивностей обслуживания заявок i-го типа ($i = \overline{1,\ I}$) одним каналом обслуживания узла j-го типа ($j = \overline{1,\ J}$) в подсистеме S;

– матрица $\mu^Z = \{\mu_{in}^Z\}$, $i = \overline{1,\ I}$, $n = \overline{1,\ N}$, состоящая из элементов μ_{in}^Z интенсивностей обслуживания заявок i-го типа ($i = \overline{1,\ I}$) одним каналом обслуживания узла n-го типа ($n = \overline{1,\ N}$) в подсистеме Z.

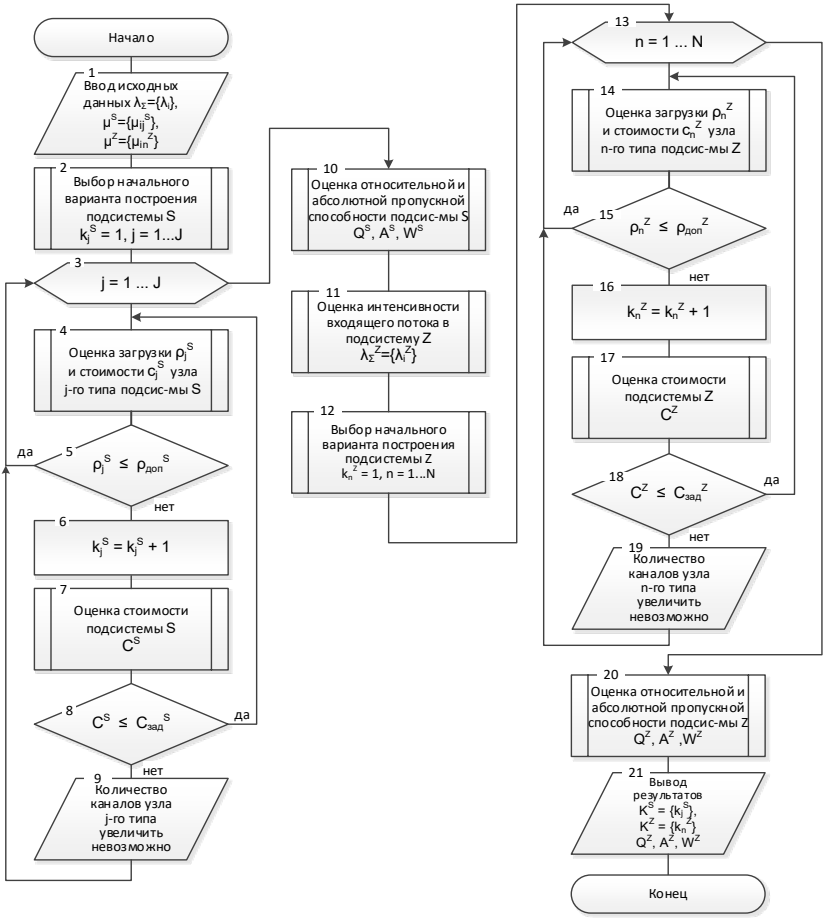


Рис. 3. Алгоритм выбора варианта построения ИУС

Алгоритм включает выполнение следующей последовательности действий:

– ввод исходных данных: $\lambda_\Sigma = \{\lambda_i\},\ i = \overline{1,\ I};\ \ \mu^S = \{\mu_{ij}^S\},\ i = \overline{1,\ I},\ j = \overline{1,\ J};\ \mu^Z = \{\mu_{in}^Z\},\ i = \overline{1,\ I},\ n = \overline{1,\ N}$ (блок 1);

– выбор начального варианта состава подсистемы S для проведения исследований: $k_j^S = 1\ ,\ \ j = \overline{1,\ J}$ (блок 2);

– организация цикла оценки загрузки и стоимости узлов j-го типа в соответствии с выражениями (32), (34) (блоки 3, 4);

– проверка выполнения условия допустимой загрузки узла j-го типа (блок 5), если условие не выполняется, то количество k_j^S каналов узла j-го типа увеличивается на 1, иначе – производится переход к выбору количества каналов узла следующего типа ($j = j+1$) (блоки 3–6);

– при увеличении количества каналов узла j-го типа производится оценка стоимости подсистемы S и проверка условия не превышения допустимых затрат на построение подсистемы S (блоки 7, 8);

– в случае выполнения условия производится переход к оценке загрузки узла j-го типа при увеличении количества каналов (блок 4), иначе – выводится сообщение о невозможности увеличения количества каналов в узле j-го типа (блок 9);

– далее производится оценка загрузки, стоимости и необходимого количества каналов для всех узлов j-го типа ($j = \overline{1, J}$) (блоки 3–9);

– по завершению цикла по типам производится оценка вероятности отказа и относительной (абсолютной) пропускной способности подсистемы S (блок 10).

Далее аналогичным образом производится анализ и выбор количества каналов подсистемы Z (блоки 11-20). По окончанию выполнения алгоритма производится вывод полученных результатов (блок 21). Полученный результат является квазиоптимальным или базовым и может быть использован в дальнейшем при более детальном исследовании ИУС.

5. Пример выбора состава обслуживающих устройств ИУС на основе её моделирования в виде сетевой СМО. Рассмотрим процесс выбора состава обслуживающих устройств ИУС, которая представлена в виде двухфазной сетевой СМО, моделирующей два этапа обслуживания заявок, поступающих на вход системы.

Пусть на вход ИУС поступает три типа заявок, характеризуемых интенсивностями: $\lambda_1 = 0{,}1$; $\lambda_2 = 0{,}3$; $\lambda_3 = 0{,}4$.

В состав ИУС входят две подсистемы: подсистема S, предназначенная для выполнения первого этапа обслуживания поступивших заявок, и подсистема Z, выполняющая обслуживание на втором этапе. В состав каждой из подсистем S и Z входят по три узла, имеющих однотипные каналы, которые характеризуются матрицей интенсивностей обслуживания заявок $\mu^S = \{\mu_{ij}^S\},\ i = \overline{1,I},\ j = \overline{1,J}$ каналами узлов в подсистеме S:

$$\mu^S = \begin{bmatrix} 0{,}3 & 0{,}4 & 0{,}5 \\ 0{,}5 & 0{,}6 & 0{,}6 \\ 0{,}6 & 0{,}7 & 0{,}8 \end{bmatrix}$$

и матрицей интенсивностей обслуживания заявок $\mu^Z = \{\mu_{in}^Z\}$, $i = \overline{1,I},\ n = \overline{1,N}$ каналами узлов в подсистеме *Z*:

$$\mu^Z = \begin{bmatrix} 0{,}2 & 0{,}3 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}5 \\ 0{,}5 & 0{,}6 & 0{,}7 \end{bmatrix}.$$

Заданы стоимости одного канала в подсистеме *S* (в у.е.):

$$c_1^S = 10;\ \ c_2^S = 20;\ \ c_3^S = 30;$$

и в подсистеме *Z*:

$$c_1^Z = 50;\ \ c_2^S = 60;\ \ c_3^S = 80,$$

а также допустимые затраты на создание подсистем *S* и *Z*:

$$C_{\text{доп}}^S = 200; \quad C_{\text{доп}}^Z = 500.$$

Кроме того, при выборе состава обслуживающих устройств учитывается допустимая загрузка соответствующих узлов подсистем *S* и *Z*:

$$\rho_{\text{доп}} = 0{,}75.$$

Требуется определить количество каналов каждого из узлов в подсистемах *S* и *Z*, при которых обеспечивается обслуживание в ИУС поступающих заявок с учетом заданных ограничений.

Решение.

1 шаг. Задание начального варианта количества каналов всех узлов подсистем *S* и *Z*:

$K_1^S = \{1, 1, 1\}$;

$K_1^Z = \{1, 1, 1\}$.

2 шаг. Задание $j = 1$. Оценка приведенной загрузки 1-го узла и суммарной стоимости подсистемы *S* при $k_1^S = 1$:

$$\rho_1^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_1^S \mu_{i1}^S} = \frac{0,1}{0,3} + \frac{0,3}{0,5} + \frac{0,4}{0,6} = 0,33 + 0,6 + 0,67 = 1,6\ ;$$

$$\left(\rho_1^S > \rho_{\text{доп}}^S\right);$$

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 10 + 20 + 30 = 60;$$

$$\left(C^S < C_{\text{доп}}^S\right).$$

3 шаг. Увеличение $k_1^S = k_1^S + 1 = 2$. Оценка приведенной загрузки 1-го узла и суммарной стоимости подсистемы S при $k_1^S = 2$:

$$\rho_1^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_1^S \mu_{i1}^S} = \frac{0,1}{0,6} + \frac{0,3}{1,0} + \frac{0,4}{1,2} = 0,17 + 0,3 + 0,33 = 0,8\ ;$$

$$\left(\rho_1^S > \rho_{\text{доп}}^S\right);$$

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 20 + 20 + 30 = 70;$$

$$\left(C^S < C_{\text{доп}}^S\right).$$

4 шаг. Увеличение $k_1^S = k_1^S + 1 = 3$. Оценка приведенной загрузки 1-го узла и суммарной стоимости подсистемы S при $k_1^S = 3$:

$$\rho_1^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_1^S \mu_{i1}^S} = \frac{0,1}{0,9} + \frac{0,3}{1,5} + \frac{0,4}{1,8} = 0,11 + 0,2 + 0,22 = 0,53\ ;$$

$$\left(\rho_1^S < \rho_{\text{доп}}^S\right);$$

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 30 + 20 + 30 = 80;$$

$\left(C^S < C^S_{\text{доп}}\right)$.

5 шаг. Задание $j = 2$. Оценка приведенной загрузки 2-го узла и суммарной стоимости подсистемы S при $k_2^S = 1$:

$$\rho_2^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_2^S \mu_{i2}^S} = \frac{0{,}1}{0{,}4} + \frac{0{,}3}{0{,}6} + \frac{0{,}4}{0{,}7} = 0{,}25 + 0{,}5 + 0{,}57 = 1{,}32\ ;$$

$\left(\rho_2^S > \rho^S_{\text{доп}}\right)$;

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 30 + 20 + 30 = 80;$$

$\left(C^S < C^S_{\text{доп}}\right)$.

6 шаг. Увеличение $k_2^S = k_2^S + 1 = 2$. Оценка приведенной загрузки 2-го узла и суммарной стоимости подсистемы S при $k_2^S = 2$:

$$\rho_2^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_2^S \mu_{i2}^S} = \frac{0{,}1}{0{,}4} + \frac{0{,}3}{1{,}2} + \frac{0{,}4}{1{,}4} = 0{,}125 + 0{,}25 + 0{,}29 = 0{,}67\ ;$$

$\left(\rho_2^S < \rho^S_{\text{доп}}\right)$;

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 30 + 40 + 30 = 100;$$

$\left(C^S < C^S_{\text{доп}}\right)$.

7 шаг. Задание $j = 3$. Оценка приведенной загрузки 3-го узла и суммарной стоимости подсистемы S при $k_3^S = 1$:

$$\rho_3^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_3^S \mu_{i3}^S} = \frac{0{,}1}{0{,}5} + \frac{0{,}3}{0{,}6} + \frac{0{,}4}{0{,}8} = 0{,}2 + 0{,}5 + 0{,}5 = 1{,}2\ ;$$

$\left(\rho_3^S > \rho^S_{\text{доп}}\right)$;

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 30 + 40 + 30 = 100;$$

$$\left(C^S < C_{\text{доп}}^S\right).$$

8 шаг. Увеличение $k_3^S = k_3^S + 1 = 2$. Оценка приведенной загрузки 3-го узла и суммарной стоимости подсистемы *S* при $k_3^S = 2$:

$$\rho_3^S = \sum_{i=1}^{I} \frac{\lambda_i}{k_3^S \mu_{i3}^S} = \frac{0{,}1}{1{,}0} + \frac{0{,}3}{1{,}2} + \frac{0{,}4}{1{,}4} = 0{,}1 + 0{,}25 + 0{,}25 = 0{,}6\ ;$$

$$\left(\rho_3^S < \rho_{\text{доп}}^S\right);$$

$$C^S = \sum_{j=1}^{J} k_j^S c_j = 30 + 40 + 60 = 130;$$

$$\left(C^S = C_{\text{доп}}^S\right).$$

9 шаг. Оценка относительной и абсолютной пропускной способности для полученного базового варианта состава обслуживающих узлов подсистемы *S* при $K_5^S = \{3, 2, 2\}$:

1) для 1-го узла подсистемы *S* (при $k_1^S = 3$):

$$P_0^{S1} = 1 / \sum_{r=1}^{k_1^S} \left(1 + \frac{\rho_1^r}{r!}\right) = 1/(1 + 0{,}53 + \frac{0{,}28}{2} + \frac{0{,}15}{6}) = 0{,}588\ ;$$

$$P_{\text{отк}}^{S1} = \frac{\rho_1^{k_1^S}}{k_1^S!} P_0^{S1} = \frac{0{,}53^3}{6} \cdot 0{,}588 = 0{,}015\ ;$$

$$Q_1^S = 1 - P_{\text{отк}}^{S1} = 1 - 0{,}022 = 0{,}985\ ;$$

2) для 2-го узла подсистемы *S* (при $k_2^S = 2$):

$$P_0^{S2} = 1/\sum_{r=1}^{k_2^S}\left(1+\frac{\rho_2^r}{r!}\right) = 1/(1+0{,}67+\frac{0{,}45}{2}) = 0{,}528\,;$$

$$P_{\text{отк}}^{S2} = \frac{\rho_1^{k_2^S}}{k_2^S!}P_0^{S2} = \frac{0{,}67^2}{2}\cdot 0{,}528 = 0{,}116\,;$$

$$Q_2^S = 1 - P_{\text{отк}}^{S2} = 1 - 0{,}116 = 0{,}884\,;$$

3) для 3-го узла подсистемы S (при $k_3^S = 2$):

$$P_0^{S3} = 1/\sum_{r=1}^{k_3^S}\left(1+\frac{\rho_3^r}{r!}\right) = 1/(1+0{,}6+\frac{0{,}36}{2}) = 0{,}561\,;$$

$$P_{\text{отк}}^{S3} = \frac{\rho_1^{k_3^S}}{k_3^S!}P_0^{S3} = \frac{0{,}6^2}{2}\cdot 0{,}561 = 0{,}101\,;$$

$$Q_3^S = 1 - P_{\text{отк}}^{S3} = 1 - 0{,}019 = 0{,}899\,;$$

4) для подсистемы S (при $K_5^S = \{3,2,2\}$):

$$P_{\text{отк}}^S = \frac{P_{\text{отк}}^{S1}\cdot P_{\text{отк}}^{S2}\cdot P_{\text{отк}}^{S3}}{P_{\text{отк}}^{S1}\cdot P_{\text{отк}}^{S2}+P_{\text{отк}}^{S1}\cdot P_{\text{отк}}^{S3}+P_{\text{отк}}^{S2}\cdot P_{\text{отк}}^{S3}} =$$

$$= \frac{0{,}015\cdot 0{,}116\cdot 0{,}101}{0{,}015\cdot 0{,}116+0{,}015\cdot 0{,}101+0{,}116\cdot 0{,}101} = 0{,}0117\,;$$

$$Q^S = 1 - P_{\text{отк}}^S = 1 - 0{,}0117 = 0{,}9883\,;$$

$$A^S = \lambda_\Sigma\cdot Q^S = 0{,}8\cdot 0{,}9883 = 0{,}7906.$$

10 шаг. Оценка интенсивности входящего потока в подсистему Z:

$$\lambda_1^Z = \lambda_1^S\cdot Q^S = 0{,}1\cdot 0{,}9883 = 0{,}09883\,;$$

$$\lambda_2^Z = \lambda_2^S\cdot Q^S = 0{,}3\cdot 0{,}9883 = 0{,}29649\,;$$

$$\lambda_3^Z = \lambda_3^S\cdot Q^S = 0{,}4\cdot 0{,}9883 = 0{,}39532\,.$$

11-20 шаги. Оценка характеристик для подсистемы Z (производится аналогичным образом) и вывод полученных результатов.

В результате оценки получены следующие значения характеристик для подсистемы Z:

1) вектор $K^Z = \{k_n^Z\},\ n = \overline{1,N}$ количества обслуживающих каналов узлов подсистемы Z:

$$K^Z = \{3, 3, 2\};$$

2) значения загрузки обслуживающих каналов узлов подсистемы Z:

$$\rho_1^Z = 0{,}68\ ;\ \rho_1^Z = 0{,}58\ ;\ \rho_1^Z = 0{,}71\ ;$$

3) затраты на создание подсистемы Z:

$$C^Z = 490 \text{ у. е.} < C_{\text{доп}}^Z\ ;$$

4) вероятности отказа в обслуживании заявок в узлах подсистемы Z:

$$P_{\text{отк}}^{Z1} = 0{,}027\ ;\quad P_{\text{отк}}^{Z2} = 0{,}018\ ;\quad P_{\text{отк}}^{Z3} = 0{,}128\ ;$$

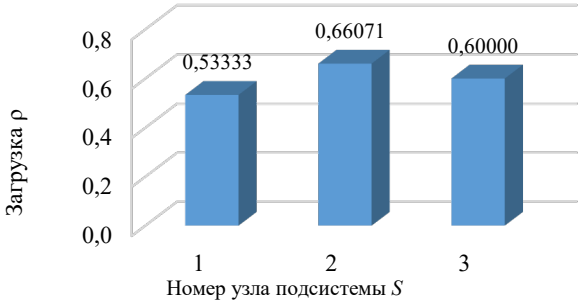
5) относительная и абсолютная пропускная способность для базового варианта состава обслуживающих узлов подсистемы Z при $K^Z = \{3, 3, 2\}$:

$$P_{\text{отк}}^Z = 0{,}0099\ ;$$

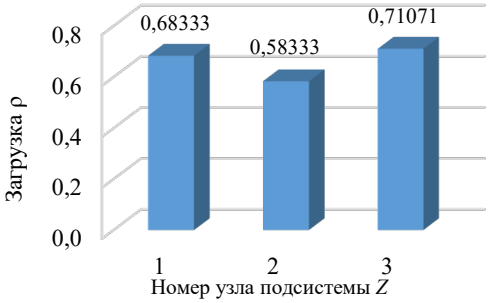
$$Q^Z = 1 - P_{\text{отк}}^Z = 1 - 0{,}0099 = 0{,}9901\ ;$$

$$A^Z = \lambda_\Sigma^Z \cdot Q^Z = 0{,}79 \cdot 0{,}9901 = 0{,}7821.$$

На рисунках 4–11 приведены зависимости загрузки, относительной пропускной способности узлов подсистем S и Z, а также показателя эффективности W системы F от её параметров. При этом на рисунке 4 показаны зависимости загрузки узлов сетевых СМО для выбранного количества каналов обслуживания в подсистемах S и Z.



а)



б)

Рис. 4. Оценка загрузки узлов сетевых СМО: а) подсистемы S; б) подсистемы Z

На рисунках 5–7 показаны зависимости относительной пропускной способности узлов сетевой СМО от количества каналов в узлах подсистемы S.

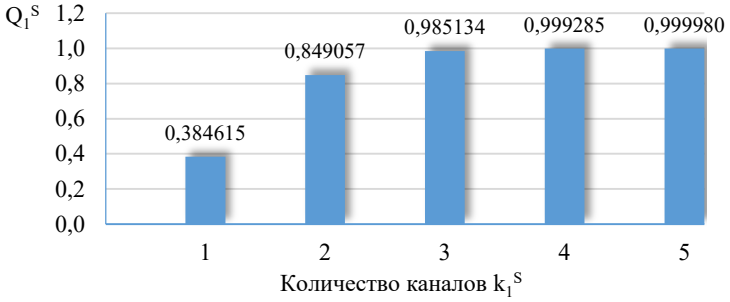


Рис. 5. Зависимость относительной пропускной способности Q_1^S от количества каналов 1-го узла подсистемы S

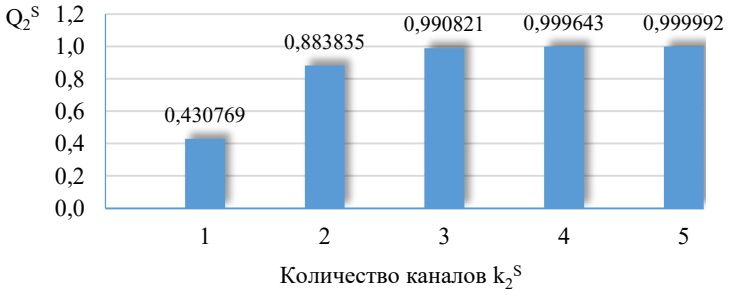


Рис. 6. Зависимость относительной пропускной способности Q_2^S от количества каналов 2-го узла подсистемы S

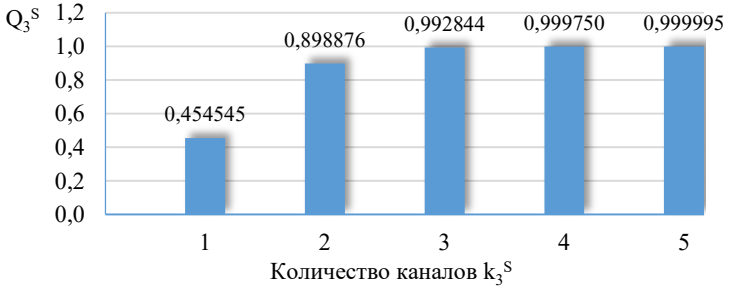


Рис. 7. Зависимость относительной пропускной способности Q_3^S от количества каналов 3-го узла подсистемы S

На рисунках 8–10 показаны зависимости относительной пропускной способности узлов сетевой СМО от количества каналов в узлах подсистемы Z.

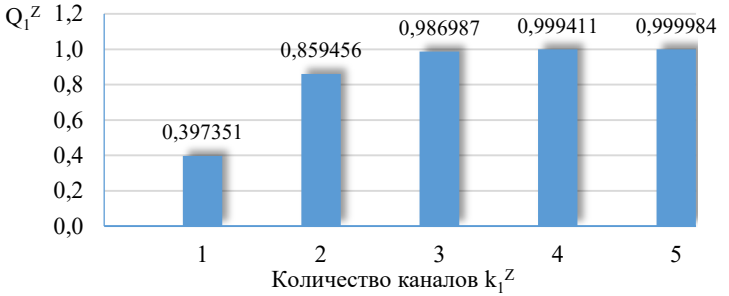


Рис. 8. Зависимость относительной пропускной способности Q_1^Z от количества каналов 1-го узла подсистемы Z

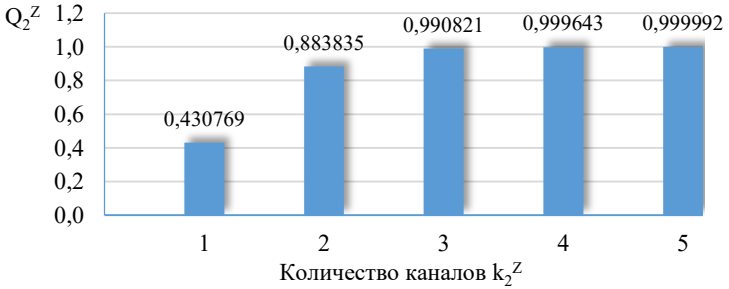


Рис. 9. Зависимость относительной пропускной способности Q_2^Z от количества каналов 2-го узла подсистемы Z

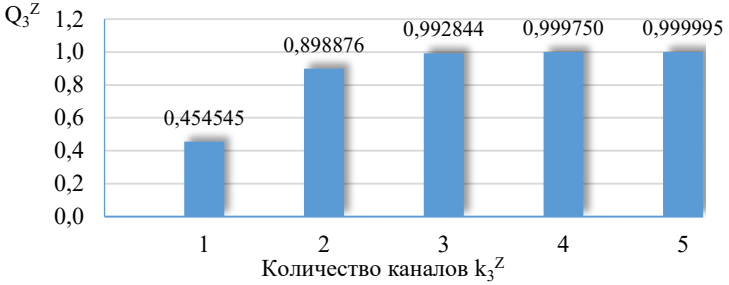


Рис. 10. Зависимость относительной пропускной способности Q_3^Z от количества каналов 3-го узла подсистемы Z

Анализ результатов моделирования показывает, что пропускная способность ИУС зависит от количества каналов в узлах и от их производительности. При этом значительное снижение загрузки происходит при увеличении количества каналов более трех или пяти в зависимости от интенсивности входного потока. Причем при загрузке менее 0,75 целесообразно повышать производительность каналов, а при загрузке более 0,75 целесообразно увеличивать количество каналов. Это объясняется резким возрастанием вероятности отказа в обслуживании заявок при большой загрузке системы. Поэтому при выборе варианта построения системы необходимо учитывать основные показатели эффективности ИУС.

На рисунке 11 показана зависимость комплексного показателя W^* эффективности системы, представляющего собой отношение прироста пропускной способности к стоимости, нормированной по отношению к максимальной стоимости варианта построения ИУС, при увеличении количества каналов в узлах сетевой СМО подсистем S и Z.

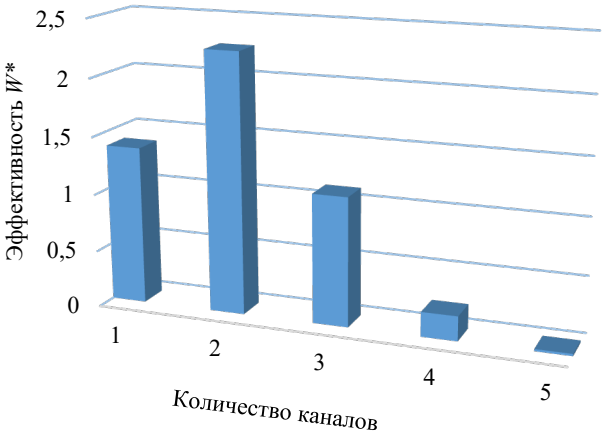


Рис. 11. Зависимость комплексного показателя эффективности от количества каналов в узлах системы

Анализ зависимости, представленной на рисунке 11, показывает, что для приведенного примера наиболее эффективным вариантом построения проектируемой ИУС является наличие в узлах обслуживания не менее 3 каналов обслуживания. Дальнейшее увеличение количества каналов (для приведенного примера – более 5) нецелесообразно, так как рост стоимости значительно превышает прирост пропускной способности системы.

Разработанный методический аппарат и предложенные принципы построения моделей ранее были использованы при выборе количества каналов обслуживания в ИУС с иерархической структурой пунктов управления [11].

Заключение. Предложенные постановка задачи и модель ИУС на основе двухфазной сетевой модели СМО являются дальнейшим совершенствованием научно-методического аппарата исследования сложных технических систем и могут быть использованы на ранних этапах проектирования ИУС. В рамках данной работы приведен практический пример выбора варианта построения ИУС и исследованы зависимости пропускной способности ИУС от количества обслуживающих каналов в узлах ИУС при обслуживании заявок и заданной информационной нагрузке. Установлено, что значительное снижение загрузки происходит при увеличении количества каналов более трех или пяти в зависимости от интенсивности входного потока. При этом увеличение пропускной

способности ИУС возможно либо путем увеличения количества каналов, либо при заданном количестве каналов путем увеличения их производительности. Эти два решения являются неравнозначными, так как, как показали исследования, при определенных условиях целесообразно повышать производительность каналов, а при других условиях целесообразно увеличивать количество каналов. Это объясняется резким возрастанием вероятности отказа в обслуживании заявок при большой загрузке системы. Поэтому при выборе варианта построения системы необходимо учитывать основные показатели эффективности ИУС.

Следующим этапом развития математических моделей подобного класса является использование в ходе исследований приоритетной обработки заявок и объединение отдельных моделей устройств ИУС на основе СМО в сетевые структуры с использованием искусственного интеллекта при формировании модели структуры ИУС.

Литература

1. Лясковский В.Л. Система поддержки принятия решений по созданию (развитию) распределенных информационно-управляющих систем организационного типа // Наукоемкие технологии в космических исследованиях Земли. 2017. Т. 9. № 6. С. 61–72.
2. Лясковский В.Л., Бреслер И.Б., Алашеев М.А. Методические и программные средства выбора решений по созданию (развитию) автоматизированных систем управления // Наукоемкие технологии в космических исследованиях Земли. 2021. Т. 13. № 3. С. 48–59.
3. Байсеитов М.Н., Ескибаев Е.Т., Избасов А.Г., Мельничук А.И., Юрков Н.К. К проблеме синтеза информационно-управляющей системы сложных технических объектов // Известия высших учебных заведений. Поволжский регион. Технические науки. 2022. № 4(64). С. 57–76.
4. Иванов С.А. Элементы информационной поддержки принятия решений при управлении лесным хозяйством // Актуальные вопросы лесного хозяйства. Материалы V международной молодежной научно-практической конференции. Санкт-Петербург: Санкт-Петербургский государственный лесотехнический университет имени С.М. Кирова, 2021. С. 138–141.
5. Фрейман В.И. К вопросу о проектировании и реализации элементов и устройств распределенных информационно-управляющих систем // Вестник ПНИПУ. Электротехника, информационные технологии, системы управления. 2019. № 30. С. 28–49.
6. Ашарина И.В., Гришин В.Ю., Сиренко В.Г. Отказоустойчивые системы управления перспективными группировками космических аппаратов как основа построения сетецентрических систем. Вопросы электромеханики. Труды ВНИИЭМ. 2023. Т. 194. № 3. С. 17–23.
7. Шумилина Н.А. Информационно-управляющая система для решения управленческих задач проектов промышленных предприятий с учетом риска отказа оборудования // Известия высших учебных заведений. Поволжский регион. Технические науки. 2018. № 2(46). С. 48–63.

8. Воеводин В.А. Модель оценки функциональной устойчивости элементов информационной инфраструктуры для условий воздействия множества компьютерных атак // Информатика и автоматизация. 2023. Т. 22. № 3. С. 691–715.
9. Козлов В.В., Лагун А.В., Харченко В.А., Коноплев М.Д. Применение иерархической системы оценивания целенаправленных процессов синтеза сложных технических систем // Известия ТулГУ. Технические науки. 2023. № 8. С. 364–374.
10. Мустафаев М.Г. Анализ эффективности функционирования и управления производственной системой предприятия // Автоматизация. Современные технологии. 2018. Т. 72. № 11. С. 499–501.
11. Онуфрей А.Ю., Разумов А.В., Какаев В.В. Метод оптимизации структуры в иерархических распределенных системах управления // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 1. С. 44–53.
12. Кротов К.В. Математическое моделирование процессов выполнения пакетов заданий в конвейерных системах с промежуточными буферами ограниченных размеров // Информатика и автоматизация. 2023. Т. 22. № 6. С. 1415–1450.
13. Клеванский Н.Н., Перетятько А.В., Леонтьев А.А., Мавзовин В.С., Воронкова И.В. Функциональная модель интегрированной системы управления учебным процессом вуза // Современные наукоемкие технологии. 2023. № 6. С. 48–55.
14. Mayr H.C., Thalheim B. The triptych of conceptual modeling // Software and Systems Modeling. 2021. vol. 20. pp. 7–24.
15. Заяц О.И., Кореневская М.М., Ильяшенко А.С., Мулюха В.А. Система массового обслуживания с абсолютным приоритетом, вероятностным выталкивающим механизмом и повторными заявками // Информатика и автоматизация. 2024. Т. 23. № 2. С. 325–351.
16. Choi B.D., Shin Y.W., Ahn W.C. Retrial queues with collision arising from unslotted CMSA/CD protocols // Queueing systems. 1992. vol. 11. pp. 335–356.
17. Кротов К.В. Математическая модель и алгоритм метода ветвей и границ для оптимизации решений по составам пакетов в многостадийных системах // Информатика и автоматизация. 2022. Т. 21. № 1. С. 5–40.
18. Андреев С.Ю., Трегубов Р.Б., Миронов А.Е. Задача выбора пропускных способностей каналов связи транспортной сети, учитывающая разбалансировку трафика различного приоритета // Труды СПИИРАН. 2020. Т. 19. № 2. С. 412–442.
19. Courtois P.J. Decomposability, instabilities, saturation in multiprogramming systems // Communications of the ACM. 1975. vol. 18. no. 7. pp. 371–377.
20. Kuhn P. Analysis of complex queuing networks by decomposition // 8th International Teletraffic Congress. 1976. pp. 236-1.
21. Qu L., Assi C., Shaban K. Network function virtualization scheduling with transmission delay optimization // IEEE/IFIP Network Operations and Management Symposium. 2016. pp. 638–644.
22. Divakaran D.M., Gurusamy M. Towards flexible guarantees in clouds: Adaptive bandwidth allocation and pricing // IEEE Transactions on Parallel and Distributed Systems. 2014. vol. 26. no. 6. pp. 1754–1764.
23. Draxler S., Karl H., Mann Z.A. Jasper: Joint optimization of scaling, placement, and routing of virtual network services // IEEE Transactions on Network and Service Management. 2018. vol. 15. no. 3. pp. 946–960.
24. Luizelli M.C., da Costa Cordeiro W.L., Buriol L.S., Gaspary L.P. A fix-and-optimize approach for efficient and large scale virtual network function placement and chaining // Computer Communications. 2017. vol. 102. pp. 67–77.

25. Андреев А.А., Шабаев А.И. Модели и методы выявления структуры локальной вычислительной сети при неполных данных // Информатика и автоматизация. 2021. Т. 20. № 1. С. 160–180.
26. Hussain T.H., Habib S.J. Capacity planning of network redesign – A case study. Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'10). 2010. pp. 52–57.
27. Zhu Z. et al. Data flow monitoring and control of LAN based on strategy. International Conference on Networking and Digital Society. 2010. vol. 2. pp. 225–228.
28. Sivakumar L, Balabaskaran J., Thulasiraman K., Arumugam S. Virtual topologies for abstraction service for IP-VPNs. 17thInternational Telecommunications Network Strategy and Planning Symposium (Networks). 2016. pp. 213–220.
29. Wang C., Huang N., Bai Y., Zhang S. A method of network topology optimization design considering application process characteristic. Modern Physics Letters B. 2018. vol. 32. no. 07. DOI: 10.1142/S0217984918500914.
30. Zhou S., Cui L., Fang C., Chai S. Research on Network Topology Discovery Algorithm for Internet of Things Based on Multi-Protocol. 10th International Conference on Modelling, Identification and Control (ICMIC). 2018. pp. 1–6. DOI: 10.1109/ICMIC.2018.8529955.
31. Пименов В.И., Пименов И.В. Анализ и визуализация данных в задачах многокритериальной оптимизации проектных решений. Информатика и автоматизация. 2022. Т. 21. № 3. С. 543–571.
32. Плескунов М.А. Теория массового обслуживания: Учебное пособие. Екатеринбург: Изд-во Урал. ун-та, 2022. 264 с.

Куракин Сергей Зосимович — канд. техн. наук, доцент, старший научный сотрудник, военный институт (научно-исследовательский), Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: теория сетевых систем массового обслуживания, численное моделирование, системный анализ, методология разработки программного обеспечения, компьютерные науки. Число научных публикаций — 125. vka@mil.ru; улица Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)237-1249.

Онуфрей Андрей Юрьевич — д-р техн. наук, профессор, работник науки и техники РФ, ведущий научный сотрудник, военный институт (научно-исследовательский), Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: прикладные задачи теории случайных процессов и теории вероятностей, методы системного анализа и построения сложных технических систем специального назначения, теория массового обслуживания. Число научных публикаций — 152. vka@mil.ru; улица Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)237-1249.

Разумов Александр Владимирович — д-р техн. наук, профессор, работник науки и техники РФ, старший научный сотрудник, военный институт (научно-исследовательский), Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: прикладные задачи построения сложных технических систем специального назначения, методы системного анализа и теории массового обслуживания. Число научных публикаций — 133. vka@mil.ru; улица Ждановская, 13, 197198, Санкт-Петербург, Россия; р.т.: +7(812)237-1249.

DOI 10.15622/ia.23.6.2

S. KURAKIN, A. ONUFREY, A. RAZUMOV

RESEARCH OF OPTIONS FOR CONSTRUCTING INFORMATION MANAGEMENT SYSTEMS BASED ON NETWORK MODELS OF QUEUING SYSTEMS

Kurakin S., Onufrey A., Razumov A. **Research of Options for Constructing Information Management Systems Based on Network Models of Queuing Systems.**

Abstract. The use of information management systems (IMSs) for the management of technical facilities is currently one of the directions for further improvement and increase in the effectiveness of the use of technical facilities in solving their target tasks. The existing modern IMSs are a set of hardware and software tools designed for collecting, processing and storing information and management. In the presence of a large amount of information and contradictory factors affecting the quality of management, making informed and timely decisions in the management process is impossible without the use of IMSs. The IMSs currently being developed are, for the most part, specialized systems and are designed to solve specific tasks. In this regard, the development and design of IMSs should be carried out taking into account the relationship with the target indicators and features of the management facilities, the results of a comprehensive analysis of information about the IMS elements in the process of functioning, and structural and algorithmic parameters that affect performance indicators. The use of mathematical models for the study of options for the construction of an IMS is the basis for the design and development of devices and subsystems of an IMS. The IMS models currently being developed make it possible to conduct research for single-stage management processes with the presence of similar service facilities in the system. At the same time, modern technical facilities and control systems are complex complexes with cyclically repeating control processes of various types of means. As a rule, such complexes have a set of parallel operating devices (control channels) that provide control of different types of objects at various stages of information processing. In this case, the structure of the IMS must be represented as a multiphase multichannel technical system in which the process of simultaneous management of several objects of various types takes place. In this regard, the purpose of the article is to develop a mathematical model of an IMS with two phases of management and the presence of an arbitrary number of serviced different types of management facilities. The basis of the model is a multiphase CFR network model with a limited waiting time for an application in the service queue. The study on the model allows choosing an option for building an IMS, in particular, choosing the optimal number of control channels for various types of objects according to the criterion of optimality and restrictions on the cost and time of management. An algorithm for selecting an option for building an IMS has been developed, and an example of calculating the number of control channels for managing three types of objects is given.

Keywords: information management system, model, mathematical modeling, network multichannel queuing system, choice of construction option, optimality criteria.

References

1. Lyaskovskij V.L. [Decision support system for the creation (development) of distributed information and control systems of an organizational type]. Naukoemkie tekhnologii v kosmicheskih issledovaniyah Zemli – High-tech in space exploration of the Earth. 2017. vol. 9. no. 6. pp. 61–72. (In Russ.).
2. Lyaskovskij V.L., Bresler I.B., Alasheev M.A. [Methodological and software tools for selecting solutions for the creation (development) of automated control systems].

Naukoemkie tekhnologii v kosmicheskih issledovaniyah Zemli – High-tech in space exploration of the Earth. 2021. vol. 13. no. 3. pp. 48–59. (In Russ.).
3. Bajseitov M.N., Eskibaev E.T., Izbasov A.G., Mel'nichuk A.I., Yurkov N.K. [On the problem of synthesis of information and control system of complex technical objects]. Izvestiya vysshih uchebnyh zavedenij. Povolzhskij region. Tekhnicheskie nauki – News of higher educational institutions. Volga region.
4. Ivanov S.A. Elementy informacionnoj podderzhki prinyatiya reshenij pri upravlenii lesnym hozyajstvom [Elements of information support for decision-making in forestry management]. Aktual'nye voprosy lesnogo hozyajstva. Materialy V mezhdunarodnoj molodezhnoj nauchno-prakticheskoj konferencii [Topical issues of forestry. Materials of the V International Youth scientific and practical conference]. St. Petersburg: St. Petersburg State Forestry Engineering University named after S.M. Kirov, 2021. pp. 138–141. (In Russ.).
5. Frejman V.I. [On the issue of design and implementation of elements and devices of distributed information and control systems]. Vestnik PNIPU. Elektrotekhnika, informacionnye tekhnologii, sistemy upravleniya – PNIPU Bulletin. Electrical Engineering, Information Technology, Control Systems. 2019. no. 30. pp. 28–49. (In Russ.).
6. Asharina I.V., Grishin V.Yu., Sirenko V.G. [Fault-tolerant control systems for promising spacecraft constellations as a basis for constructing network-centric systems]. Voprosy elektromekhaniki. Trudy VNIIEM – Questions of Electromechanics. Proceedings of VNIIEM. 2023. vol. 194. no. 3. pp. 17–23. (In Russ.).
7. Shumilina N.A. [Information and management system for solving management problems of industrial enterprise projects taking into account the risk of equipment failure]. Izvestiya vysshih uchebnyh zavedenij. Povolzhskij region. Tekhnicheskie nauki – News of higher educational institutions. Volga region. Technical sciences. 2018. no. 2(46). pp. 48–63. (In Russ.).
8. Voevodin V.A. [A model for assessing the functional stability of information infrastructure elements for conditions of exposure to multiple computer attacks]. Informatika i avtomatizaciya – Informatics and Automation. 2023. vol. 22. no. 3. pp. 691–715. (In Russ.).
9. Kozlov V.V., Lagun A.V., Harchenko V.A., Konoplev M.D. [Application of a hierarchical system for evaluating targeted processes of synthesis of complex technical systems]. Izvestiya TulGU. Tekhnicheskie nauki – Tula State University News. Technical Sciences. 2023. no. 8. pp. 364–374. (In Russ.).
10. Mustafaev M.G. [Analysis of the efficiency of functioning and management of the enterprise production system] Avtomatizaciya. Sovremennye tekhnologii – Automation. Modern technologies. 2018. vol. 72. no. 11. pp. 499–501. (In Russ.).
11. Onufrej A.Yu., Razumov A.V., Kakaev V.V. [Method of structure optimization in hierarchical distributed control systems]. Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki – Scientific and technical journal of information technologies, mechanics and optics. 2023. vol. 23. no. 1. pp. 44–53. (In Russ.).
12. Krotov K.V. [Mathematical modeling of the processes of completing task packages in conveyor systems with intermediate buffers of limited size]. Informatika i avtomatizaciya – Informatics and Automation. 2023. vol. 22. no. 6. pp. 1415–1450. (In Russ.).
13. Klevanskij N.N., Peretyat'ko A.V., Leont'ev A.A., Mavzovin V.S., Voronkova I.V. [Functional model of the integrated system of management of the educational process of the university]. Sovremennye naukoemkie tekhnologii – Modern high-tech technologies. 2023. no. 6. pp. 48–55. (In Russ.).

14. Mayr H.C., Thalheim B. The triptych of conceptual modeling. Software and Systems Modeling. 2021. vol. 20. pp. 7–24.
15. Zayac O.I., Korenevskaya M.M., Il'yashenko A.S., Mulyuha V.A. [Queuing system with absolute priority, probabilistic ejection mechanism and repeated]. Informatika i avtomatizaciya – Informatics and Automation. 2024. vol. 23. no. 2. pp. 325–351. (In Russ.).
16. Choi B.D., Shin Y.W., Ahn W.C. Retrial queues with collision arising from unslotted CMSA/CD protocols. Queueing systems. 1992. vol. 11. pp. 335–356.
17. Krotov K.V. A mathematical model and algorithm of the branch and boundary method for optimizing package composition solutions in multistage systems. Informatika i avtomatizaciya – Informatics and Automation. 2022. vol. 21. no. 1. pp. 5–40. (In Russ.).
18. Andreev S.Yu., Tregubov R.B., Mironov A.E. [The task of selecting the bandwidth of communication channels of the transport network, taking into account the imbalance of traffic of different priorities]. Trudy SPIIRAN – Proceedings of SPIIRAN 2020. vol. 19. no. 2. pp. 412–442. (In Russ.).
19. Courtois P.J. Decomposability, instabilities, saturation in multiprogramming systems. Communications of the ACM. 1975. vol. 18. no. 7. pp. 371–377.
20. Kuhn P. Analysis of complex queuing networks by decomposition. 8th International Teletraffic Congress. 1976. pp. 236-1.
21. Qu L., Assi C., Shaban K. Network function virtualization scheduling with transmission delay optimization. IEEE/IFIP Network Operations and Management Symposium. 2016. pp. 638–644.
22. Divakaran D.M., Gurusamy M. Towards flexible guarantees in clouds: Adaptive bandwidth allocation and pricing. IEEE Transactions on Parallel and Distributed Systems. 2014. vol. 26. no. 6. pp. 1754–1764.
23. Draxler S., Karl H., Mann Z.A. Jasper: Joint optimization of scaling, placement, and routing of virtual network services. IEEE Transactions on Network and Service Management. 2018. vol. 15. no. 3. pp. 946–960.
24. Luizelli M.C., da Costa Cordeiro W.L., Buriol L.S., Gaspary L.P. A fix-and-optimize approach for efficient and large scale virtual network function placement and chaining. Computer Communications. 2017. vol. 102. pp. 67–77.
25. Andreev A., Shabaev A. [Models and Methods for Discovery of Local Area Network Topology with Incomplete Data]. Informatika i avtomatizaciya – Informatics and Automation. 2021. vol. 20(1). pp. 160–180. (In Russ.).
26. Hussain T.H., Habib S.J. Capacity planning of network redesign – A case study. Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'10). 2010. pp. 52–57.
27. Zhu Z. et al. Data flow monitoring and control of LAN based on strategy. International Conference on Networking and Digital Society. 2010. vol. 2. pp. 225–228.
28. Sivakumar L, Balabaskaran J., Thulasiraman K., Arumugam S. Virtual topologies for abstraction service for IP-VPNs. 17thInternational Telecommunications Network Strategy and Planning Symposium (Networks). 2016. pp. 213–220.
29. Wang C., Huang N., Bai Y., Zhang S. A method of network topology optimization design considering application process characteristic. Modern Physics Letters B. 2018. vol. 32. no. 07. DOI: 10.1142/S0217984918500914.
30. Zhou S., Cui L., Fang C., Chai S. Research on Network Topology Discovery Algorithm for Internet of Things Based on Multi-Protocol. 10th International Conference on Modelling, Identification and Control (ICMIC). 2018. pp. 1–6. DOI: 10.1109/ICMIC.2018.8529955.

31. Pimenov V., Pimenov I. [Data Analysis and Visualization in the Tasks of the Project Solutions Multicriteria Optimization]. Informatika i avtomatizaciya – Informatics and Automation. 2022. vol. 21(3). pp. 543–571. (In Russ.).
32. Pleskunov M.A. Teoriya massovogo obsluzhivaniya: Uchebnoe posobie [Queuing Theory: A textbook.]. Ekaterinburg: Izd-vo Ural. un-ta, 2022. 264 p. (In Russ.).

Kurakin Sergey — Ph.D., Associate Professor, Senior researcher, Military institute (scientific research), The Military Space Academy named after A.F. Mozhaisky. Research interests: theory of network queuing systems, numerical modeling, software development methodologies, computer science. The number of publications — 125. vka@mil.ru; 13, Zhdanovskaya St., 197198, St. Petersburg, Russia; office phone: +7(812)237-1249.

Onufrey Andrey — Ph.D., Dr.Sci., Professor, Leading researcher, Military institute (scientific research), The Military Space Academy named after A.F. Mozhaisky. Research interests: applied problems of the theory of random processes and probability theory, methods of system analysis and construction of complex technical systems for special purposes, queuing theory. The number of publications — 152. vka@mil.ru; 13, Zhdanovskaya St., 197198, St. Petersburg, Russia; office phone: +7(812)237-1249.

Razumov Alexander — Ph.D., Dr.Sci., Professor, Senior researcher, Military institute (scientific research), The Military Space Academy named after A.F. Mozhaisky. Research interests: applied problems of building complex technical systems for special purposes, methods of system analysis and queuing theory. The number of publications — 133. vka@mil.ru; 13, Zhdanovskaya St., 197198, St. Petersburg, Russia; office phone: +7(812)237-1249.

УДК 519.178 **DOI** 10.15622/ia.23.6.3

В.Н. Куделя

РЕШЕНИЕ ЗАДАЧ ПЕРЕБОРА ПУТЕЙ В СЛОЖНЫХ ГРАФАХ

Куделя В.Н. **Решение задач перебора путей в сложных графах.**

Аннотация. Моделирование различных систем связано с перебором значений параметров элементов структуры и учетом всех характеристик функционирования и взаимодействия компонентов для нахождения определенного набора решений, определяющих конфигурацию системы. Такие задачи относятся к задачам переборного типа и подразумевают, что некоторое количество очередных решений из этого набора получается из предыдущего решения в определенном порядке. Известно, что достаточно большое количество задач переборного типа решается только методами полного перебора и других методов для их точного решения пока не существует. В статье представлен новый метод перебора путей в графе – метод трансформации узлов-графов. По предварительной оценке, предложенный метод, в отличие от существующих, позволяет значительно быстрее осуществлять поиск всех простых путей в ориентированном графе произвольной структуры. В известных методах перебора в графе (Breadth First Search и Depth First Search) объектом перебора является путь. Всё количество таких путей в графе определяет размер пространства перебора. Основная идея метода трансформации узлов-графов заключается в значительном уменьшении размера пространства перебора за счет укрупнения объектов перебора. Укрупнение объектов перебора осуществляется кластеризацией путей в комбинаторные объекты, объединяющие по определенному регламенту некоторое множество путей одинаковой длины. Такие комбинаторные объекты названы узлами-графами. Узел-граф относится к центрально-периферическим комбинаторным объектам и для перебора всех путей в графе разработаны специфические операции преобразования узлов-графов, которые позволяют найти следующие пути на основе предыдущих. Метод может использоваться как базовый инструментарий для уменьшения размерности пространства поиска решений NP-полных задач, сохраняя универсальность и точность перебора.

Ключевые слова: граф, путь, гамильтонов путь, перебор, комбинаторный взрыв, NP-полные задачи, кластеризация, узел-граф.

1. Введение. С бурным развитием информационных технологий теория графов резко расширило сферу своего применения в различных областях знаний. Графы получили исключительную популярность при моделировании таких систем, как распределенные бухгалтерские книги типа направленных ациклических графов [1], социальные сети [2], транспортные системы [3, 4], системы связи [5], системы комбинаторных аукционов [6], финансовые системы [7] и т.д.

Моделирование всех компонентов системы, таких как структура, регламенты, алгоритмы функционирования и других, обычно сводится к решению комплекса задач теории графов, которые можно условно объединить в несколько групп:

– задачи анализа связности: оценка наличия связи между какими-либо элементами графа;

– задачи упорядочения вершин или задачи кластеризации: выделение неких скоплений в графе. Например, независимые множества вершин или множества вершин с одинаковыми характеристиками и т.д.;

– задачи размещения центров: выделение каких-либо элементов графа с уникальными свойствами, например, минимаксное множество центров, множество центров с минимальной суммой и т.д.;

– задачи о маршрутах: путь, гамильтонов путь, независимые пути и т.д.;

– потоковые задачи: задача о максимальном потоке, поток в сети с ограничениями, распараллеливание потока задач, распределение потоков данных и т.д.

Однако известно [8], что наибольшее количество практически важных задач теории графов входят в список NP-полных задач. Кроме того, любая NP-полная задача может быть решена методом полного перебора [9 – 11], например, задача перечисления путей. Количество путей длиной от $p=1$ до $p=v-1$ в полном ориентированном графе составляет $ev!$ ($e=2,71828...$ – число Эйлера, $v=|V|$ – мощность множества вершин V графа $G[V,E]$). Из-за такого количества результатов (размера пространства перебора) задача перечисления путей относится к задачам с комбинаторным взрывом [12], что и определяет высокий уровень сложности большинства NP-полных задач.

Наиболее известными являются два метода перебора путей в графе – поиск в ширину (Breadth First Search, BFS) и поиск в глубину (Depth First Search, DFS). А метод возведения в степень матрицы смежности такого распространения не получил [13]. Методы BFS и DFS исследованы в достаточном объеме и их использование при решении многих задач переборного типа приводит к комбинаторному взрыву. Комбинаторный взрыв происходит, когда создается громадное количество возможных комбинаций за счет увеличения количества объектов перебора. Для обхода проблемы комбинаторного взрыва разрабатывают специальные методы решения. Чаще всего применяют эвристические алгоритмы, которые основываются на предположениях о наиболее перспективных ветвях процесса перебора, остальные ветви исключаются. Исходя из необходимости преодоления комбинаторного взрыва, автор разработал и исследовал несколько методов перебора путей в графе и способов кластеризации результатов перебора:

1. Разработаны следующие методы перебора:

– метод транспонирования [14], позволяющий решать задачу поиска всех путей в графе, так же как и методы BFS и DFS, перебирая пути длины p для поиска путей длины $p = p+1$;

– инверсный метод [15], дающий возможность перебирать пути длиной p для поиска путей длиной $p+k$, где $k = \{1,2,3...\}$ – размер прироста (шага) длины пути. Пространство перебора сократилось и при размере шага $k = 2$ составило $\frac{8\sqrt{3}}{9}v!$, где $\frac{8\sqrt{3}}{9}$ – квадратная ледяная постоянная *Либа*.

2. Исследование разработанных методов позволило подтвердить некоторые причины комбинаторного взрыва и сформулировать новые положения по ускорению алгоритмов перебора:

– вычислительная сложность решения задач поиска всех путей в графе обусловлена тем, что чем больше длина p отыскиваемых путей, тем больше их возможное количество;

– количество искомых путей длиной $p+1$, находится в зависимости вида $O_{p+1} = O_p(v-p-1)$ от количества путей длиной p;

– сложность решения NP-полных задач может быть снижена, например при использовании инверсного метода, для решения задачи поиска всех гамильтоновых путей в графе. При этом точность полученных результатов сохраняется;

– один из основных способов значительного снижения эффекта комбинаторного взрыва – это сокращение пространства перебора, но не с помощью эвристик [16 – 20], дающих вероятные результаты, а за счет объединения путей в группы (кластеры), то есть объединить в более крупные объекты перебора [14, 15];

– дополнительно сократить пространство перебора и, как следствие, ускорить алгоритм перебора можно при увеличении, где это допустимо, размера шага $k = \{2,4,6...\}$.

В данной статье представляется новый метод перебора путей в графе – метод трансформации узлов-графов. Для демонстрации метода трансформации узлов-графов автор выбрал задачу перечисления гамильтоновых путей в графе [18]. Постановка задачи: какова мощность множества гамильтоновых путей в графе $G[V,E]$.

2. Общие положения. Автор использовал следующие термины.

Путь – последовательность ребер $e_1, e_2, \ldots, e_k$, в которой конец одного ребра является началом следующего ребра. Начало первого ребра называется началом пути, конец последнего ребра – концом пути.

Простой путь – путь, который проходит через каждую вершину не более одного раза.

Длина пути p – число ребер (дуг), которые его образуют.

Гамильтонов путь – это простой путь, проходящий через все вершины графа.

Окрестность вершины v_k в графе – подграф графа $G[V,E]$, состоящий из всех вершин, сопряжённых v_k, и всех рёбер, соединяющих две такие вершины.

Узел-граф – подграф графа $G[V,E]$, состоящий из образующей подмножество U последовательности вершин $u_1, u_2, \ldots, u_i$ ($i = \{1,2,3,...\}$, $i \le v-2$), из подмножества S вершин, не входящих в U и имеющих исходящую дугу в u_1, и из подмножества T вершин, не входящих в U и имеющих исходящую дугу из u_i, а также всех дуг, соединяющих такие вершины с первой u_1 и последней u_i из подмножества U.

Схема кластеризации – схема группирования комбинаторных объектов, в нашем случае путей.

В дальнейшем, если не оговаривается иное, под графом понимается ориентированный граф без петель. В графе ищутся простые пути, далее – путь. Для обозначения вершины будем использовать одну цифру, а цифры через запятую – это дуга, соединяющая смежные вершины. Точка с запятой разделяет числа, обозначающие вершины, не имеющие признака смежности в подмножестве.

На изображениях графа, для того чтобы не загромождать рисунок, ненаправленными ребрами будут обозначаться две противоположно направленные дуги.

3. Метод трансформации узлов-графов

3.1. Представление графа. Новое представление графа будет далее рассмотрено на основе примера графа, изображенного на рисунке 1.

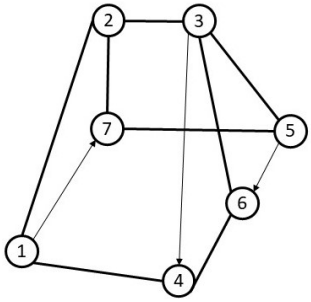


Рис. 1. Граф $G[V,E]$

Выделим в графе $G[V,E]$ одну из окрестностей, например для вершины 5, смотри рисунке 2.

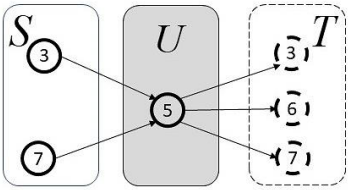


Рис. 2. Узел-граф

Вершины из окрестности (сплошной контур), имеющие исходящую дугу, направленную к вершине с серой заливкой, размещаем слева от вершины с серой заливкой. Это подмножество вершин назовем истоком S вершины с серой заливкой. Вершины из окрестности (штрих-контур), имеющие входящую дугу, направленную из вершины с серой заливкой, размещаем справа от вершины с серой заливкой. Подмножество вершин, имеющих штрих-контур, будем называть стоком T вершины с серой заливкой.

Будем считать, что путь от какой-либо вершины в себя существует всегда – это путь длиной $p=0$. Исходя из этого, примем, что область с серой заливкой тоже является подмножеством – телом U узла-графа. Кластер подмножеств S, U, T и является узлом-графом. В отличие от окрестности вершины, узел-граф связан с местом сосредоточения путей, таких как перекресток, железнодорожный узел, центр города, где пересекаются пути. То есть узел-граф – подграф

графа $G[V,E]$, состоящий из образующей подмножество U последовательности вершин $u_1, u_2, \ldots, u_i$ ($i = \{1,2,3,...\}$, $i \le v-2$), из подмножества S вершин, не входящих в U и имеющих исходящую дугу в u_1, и из подмножества T вершин, не входящих в U и имеющих исходящую дугу из u_i, а также всех дуг, соединяющих такие вершины с первой u_1 и последней u_i из подмножества U.

Символьное представление узла-графа имеет вид ${}^{(q)}_{[S]}U^{(w)}_{[T]}$. q и w будут определены далее. Схема кластеризации путей в узел-граф представлена в таблице 1.

Таблица 1. Схема кластеризации путей в узел-граф

№ п/п	Отношения в подмножествах и между подмножествами
1.	Тело узла-графа U – подмножество последовательности смежных вершин. $U \subset V$, $u = \lvert U \rvert$ – мощность множества U, $u = \{1,2,3,...\}$, $u \le v-2$
2.	Исток узла-графа S – подмножество несвязанных вершин. $S \subset V$, $s = \lvert S \rvert$ – мощность множества S, $s = \{0,1,2,3,...\}$, $s \le (v-1)$
3.	Исток узла-графа размещается слева от тела. Вершина из истока S имеет только исходящую дугу, направленную к первой вершине из тела U, $S \cap U = \varnothing$
4.	Сток узла-графа T – подмножество несвязанных вершин. $T \subset V$, $t = \lvert T \rvert$ – мощность множества T, $t = \{0,1,2,3,...\}$, $t \le (v-1)$
5.	Сток размещается справа от тела. Вершина из стока T имеет только входящую дугу, направленную от последней вершины из тела U, $T \cap U = \varnothing$

Для удобства демонстрации операций с узлами-графами, элементы подмножеств S и T могут располагаться либо горизонтально через точку с запятой, либо – вертикально без разделительных знаков.

Параметры q и w характеризуют узел-граф. Данные параметры являются дополнительными, их наличие и количество необязательно. Это могут быть:

– внутренние параметры узла-графа: количество элементов, вес, длина, стоимость и другие;

– параметры моделируемого объекта: параметры взаимодействия со средой, в которой функционирует объект; технические; химические; биоинформационные; генетические и т.д. характеристики объекта.

Совокупность всех узлов-графов, у которых мощность множества $u=1$, полностью описывает граф $G[V,E]$, то есть является способом представления графа, эквивалентным всем существующим способам, так как на основе любого представления можно получить все остальные. Каждый узел-граф, у которого мощность множества $u=1$, объединяет все направленные пути длиной $p=2$, проходящие через тело U. Все пути в узле-графе направлены от истока S к стоку T.

Тогда граф $G[V,E]$ (рисунок 1) задается следующей совокупностью узлов-графов:

$$^{(4)}_{[2;4]}1^{(1)}_{[2;4;7]} \quad ^{(6)}_{[1;3;7]}2^{(2)}_{[1;3;7]} \quad ^{(9)}_{[2;5;6]}3^{(3)}_{[2;4;5;6]} \quad ^{(4)}_{[1;3;6]}4^{(4)}_{[1;6]} \quad ^{(4)}_{[3;7]}5^{(5)}_{[3;6;7]} \quad ^{(4)}_{[3;4;5]}6^{(6)}_{[3;4]} \quad ^{(4)}_{[1;2;5]}7^{(7)}_{[2;5]}$$

Для вершин из подмножеств S и T признак смежности не указывается, даже если они являются смежными в графе.

Здесь и далее q – число путей, проходящих через тело U, а $w=\{1,2,3,...\}$ – номер узла-графа в совокупности.

Количество путей длины p в графе можно рассчитать по формуле $O_p=\dfrac{v!}{(v-p-1)!}$. Тогда в полном ориентированном графе, состоящем из семи вершин, имеется 210 путей (объектов перебора) длиной $p=2$. При этом каждый путь содержит 3 вершины (элемента). То есть все пути включают 630 элементов. Тот же граф, представленный узлами-графами, включает 7 объектов перебора. Каждый узел-граф может содержать максимум 13 элементов. Тогда представление графа совокупностью узлов-графов содержит 91 элемент, что почти в 7 раз меньше. С увеличением числа вершин в графе степень отличия возрастает.

Кластеризация элементов (вершин, дуг) множеств V и E в группы подмножеств S, U и T позволяет представить граф в более компактной форме – в виде совокупности узлов-графов. Очевидно, что количество объектов перебора длиной p для поиска объектов длиной $p+k$ уменьшилось. Таким образом, пространство перебора за счет укрупнения объектов перебора сократилось.

Кластеризация может производиться и по другим основаниям, а узел-граф может создаваться и для других объектов перебора. Внутри подмножеств S, U и T также могут формироваться группы объектов.

3.2. Операции трансформации узла-графа. Здесь представлены операции трансформации узла-графа для поиска всех путей в совокупности узлов-графов. Операции трансформации узла-графа позволяют найти пути длиной $p+k$ на основе путей длиной p, то есть осуществить исчерпывающий поиск путей в графе.

При описании и демонстрации операций трансформации узла-графа подразумевается, что параметр w, номер узла-графа, относится ко всем составляющим узел-графа. Операции трансформации узла-графа выполняются для комбинаторных объектов, таких как вершина, дуга, узел-граф, находящихся в отношениях смежности. Узлы-графы ${}^{(q)}_{[S]}U^{(1)}_{[T]}$ и ${}^{(q)}_{[S]}U^{(2)}_{[T]}$ являются смежными, если вершина $i \in T^{(1)}$ и $i \in S^{(2)}$ или $i \in T^{(2)}$ и $i \in S^{(1)}$.

Узел-граф относится к центрально-периферическим комбинаторным объектам, и процесс поиска путей в графе ассоциируется с последовательным поглощением центром (телом) U периферийных элементов (смежных узлов-графов, вершин из S и T).

Любой узел-граф ($u \geq 1$) разрешено использовать в операциях неоднократно, например, для увеличения шага k. Для перебора в графе каждый узел-граф выполняет функции то центра, то периферии. Любая операция трансформации может выполняться совершенно независимо для различных узлов-графов, то есть обеспечивается параллельный процесс получения результатов перебора.

Узел-граф, имеющий $u > 1$, существует при условии:

– если $S^{(w)} \neq \varnothing \wedge T^{(w)} \neq \varnothing$, путь обязательно где-то начинается и куда-то приводит;

– если $s = t = 1$ и $S^{(w)} \neq T^{(w)}$, циклы исключаются.

3.2.1. Добавление вершин. Регламент добавления вершин в узел-граф представлен в таблице 2.

Таблица 2. Регламент добавления вершин в узел-граф

№ п/п	Регламент	Операции
1.	$d = \max w + 1$	$+$ – сложения
2.	Если $i \in S^{(w)}$, $j \in T^{(w)}$ и $i \neq j$, то $U^{(w=d)} = i + U^{(w)} + j$	$+$ – добавления вершины
3.	$S^{(w=d)} = S^{(w=i)} \setminus (S^{(w=i)} \cap U^{(w=d)})$	$\setminus$ – разность множеств
4.	$T^{(w=d)} = T^{(w=j)} \setminus (T^{(w=j)} \cap U^{(w=d)})$	$\cap$ – пересечения множеств
5.	Проверка условий существования узла-графа $\begin{pmatrix} y^{(w=d)} \\ \left[S^{(w=d)}\right] \end{pmatrix} U^{(w=d)}_{\left[T^{(w=d)}\right]}$	удаление узла-графа, если условие не выполняется
6.	Пересчет значения параметра q узла-графа	

Существуют следующие ограничения на операцию добавления вершин:

– мощность множества вершин графа v не менее 5;

– операция выполняется для узлов-графов, у которых $S^{(w)} \neq \varnothing$ и $T^{(w)} \neq \varnothing$.

Очевидно, что множество путей длиной $p+2$ получается из множества путей длиной p, то есть размер шага $k = 2$.

Добавление одной вершины с одной из сторон является частным случаем добавления вершин в узел-граф.

Алгоритм добавления вершин проиллюстрируем на узле-графе № 3 для графа, представленного на рисунке 1. В ячейки $f_{i,j}$ таблицы F размером $v \times v$, у которых $i \in S^{(3)}$, $j \in T^{(3)}$ и $i \neq j$, вписывается узел-граф $\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$.

Номера вершин стока j

Номера вершин истока i	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0
3	0	0	0	0	0	0	0
4	0		0	0			0
5	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0
6	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	$\begin{pmatrix} \vartheta^{(3)} \\ \left[2;5;6^{(3)}\right] \end{pmatrix} 3^{(3)}_{\left[2;4;5;6^{(3)}\right]}$	0	0
7	0	0	0	0	0	0	0

Далее выполняем регламент добавления вершин для узла-графа № 3. Например, для ячейки $f_{2,4}$, определим:

1. $d = \max w + 1 = 7 + 1 = 8$;
2. $U^{(8)} = i + U^{(3)} + j = 2 + 3 + 4 = 2,3,4$;
3. $S^{(8)} = S^{(2)} \setminus (S^{(2)} \cap U^{(8)}) = 1;3;7 \setminus (1;3;7 \cap 2;3;4) = 1;3;7 \setminus 3 = 1;7$;
4. $T^{(8)} = T^{(4)} \setminus (T^{(4)} \cap U^{(8)}) = 1;6 \setminus (1;6 \cap 2;3;4) = 1;6 \setminus \varnothing = 1;6$;
5. $S^{(8)} \neq \varnothing$, $T^{(8)} \neq \varnothing$, значит, $\left[{}^{(q^{(8)})}_{S^{(8)}}\right] U^{(8)}_{\left[T^{(8)}\right]}$ существует.
6. $q^{(8)} = 5$.

В результате трансформации узла-графа № 3 имеем.

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	0	$\left[{}^{(3^{(8)})}_{1;7^{(8)}}\right] 2,3,4^{(8)}_{\left[1;6^{(8)}\right]}$	$\left[{}^{(3^{(9)})}_{1;7^{(9)}}\right] 2,3,5^{(9)}_{\left[6;7^{(9)}\right]}$	$\left[{}^{(2^{(10)})}_{1;7^{(10)}}\right] 2,3,6^{(10)}_{\left[4^{(10)}\right]}$	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	$\left[{}^{(1^{(11)})}_{7^{(11)}}\right] 5,3,2^{(11)}_{\left[1;7^{(11)}\right]}$	0	$\left[{}^{(2^{(12)})}_{7^{(12)}}\right] 5,3,4^{(12)}_{\left[1;6^{(12)}\right]}$	0	$\left[{}^{(1^{(13)})}_{7^{(13)}}\right] 5,3,6^{(13)}_{\left[4^{(13)}\right]}$	0
6	0	$\left[{}^{(4^{(14)})}_{4;5^{(14)}}\right] 6,3,2^{(14)}_{\left[1;7^{(14)}\right]}$	0	$\left[{}^{(1^{(15)})}_{5^{(15)}}\right] 6,3,4^{(15)}_{\left[1^{(15)}\right]}$	$\left[{}^{(2^{(16)})}_{4^{(16)}}\right] 6,3,5^{(16)}_{\left[7^{(16)}\right]}$	0	0
7	0	0	0	0	0	0	0

Так как $p \neq v - 1$, продолжим трансформацию для получения гамильтоновых путей, например для узла-графа № 9, и получим.

	1	2	3	4	5	6	7
1	0	0	0	0	0	$\left[{}^{(\varnothing^{(17)})}_{4^{(17)}}\right] 1,2,3,5,6^{(17)}_{\left[4^{(17)}\right]}$	$\left[{}^{(\varnothing^{(17)})}_{4^{(17)}}\right] 1,2,3,5,7^{(17)}_{\left[\varnothing^{(17)}\right]}$
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0		
6	0	0	0	0	0	0	0
7	0	0	0	0	0	$\left[{}^{(3^{(17)})}_{1^{(17)}}\right] 7,2,3,5,6^{(17)}_{\left[4^{(17)}\right]}$	0

3.2.2. Добавление узлов-графов. В отличие от операции добавления вершин, операция добавления узлов-графов предполагает

добавление (поглощение) как вершин из S и T, так и подмножеств U в тройке смежных узлов-графов. При этом порядок элементов в подмножествах U сохраняется. Операция добавления будет обозначаться символом $\|$.

Пусть выбраны три узла-графа $\begin{matrix}\left(q^{(1)}\right)\\ \left[S^{(1)}\right]\end{matrix}U^{(1)}_{\left[T^{(1)}\right]}$, $\begin{matrix}\left(q^{(2)}\right)\\ \left[S^{(2)}\right]\end{matrix}U^{(2)}_{\left[T^{(2)}\right]}$ и $\begin{matrix}\left(q^{(3)}\right)\\ \left[S^{(3)}\right]\end{matrix}U^{(3)}_{\left[T^{(3)}\right]}$ графа $G[V,E]$.

Добавлением узлов-графов называется узел-граф $\begin{matrix}\left(q^{(w=d)}\right)\\ \left[S^{(w=d)}\right]\end{matrix}U^{(w=d)}_{\left[T^{(w=d)}\right]} = \begin{matrix}\left(q^{(1)}\right)\\ \left[S^{(1)}\right]\end{matrix}U^{(1)}_{\left[T^{(1)}\right]} \| \begin{matrix}\left(q^{(2)}\right)\\ \left[S^{(2)}\right]\end{matrix}U^{(2)}_{\left[T^{(2)}\right]} \| \begin{matrix}\left(q^{(3)}\right)\\ \left[S^{(3)}\right]\end{matrix}U^{(3)}_{\left[T^{(3)}\right]}$, полученный по регламенту, представленному в таблице 3.

Таблица 3. Регламент добавления узлов-графов

№ п/п	Регламент	Операции
1.	$d = \max w + 1$	$+$ – сложения
2.	Если $i \in T^{(1)}$, $i \in S^{(2)}$, $j \in T^{(2)}$, $j \in S^{(3)}$, то $U^{(w=d)} = U^{(1)} \|(i + U^{(2)} + j) \|U^{(3)}$	$+$ – добавления вершин $\|$ – добавления тела узла-графа
3.	$S^{(w=d)} = S^{(1)} \setminus (S^{(1)} \cap U^{(w=d)})$	$\setminus$ – разность множеств $\cap$ – пересечение множеств
4.	$T^{(w=d)} = T^{(3)} \setminus (T^{(3)} \cap U^{(w=d)})$	
5.	Проверка условий существования узла-графа $\begin{matrix}\left(y^{(w=d)}\right)\\ \left[S^{(w=d)}\right]\end{matrix}U^{(w=d)}_{\left[T^{(w=d)}\right]}$.	удаление узла-графа, если условие не выполняется
6.	Пересчет значения параметра q узла-графа	

Ограничения на операцию добавления узлов-графов:

– мощность множества вершин графа v не менее 7;

– операция выполняется для узлов-графов, у которых $S^{(1)} \neq \varnothing \ \wedge \ T^{(1)} \neq \varnothing \ \wedge \ S^{(2)} \neq \varnothing \ \wedge \ T^{(2)} \neq \varnothing \ \wedge \ S^{(3)} \neq \varnothing \ \wedge \ T^{(3)} \neq \varnothing$, а также $U^{(1)} \cap U^{(2)} = \varnothing$, $U^{(1)} \cap U^{(3)} = \varnothing$, $U^{(2)} \cap U^{(3)} = \varnothing$.

Очевидно, что множество путей длиной $p+4$ получается из множества путей длины p, то есть $k = 4$. Добавление двух узлов-графов с одной из сторон является частным случаем добавления узлов-графов.

Алгоритм операции добавления узлов-графов проиллюстрируем на узлах-графах № 1, 6 и 5 для графа, представленного на рисунке 1:

Шаг 1. В ячейки $f_{i,j}$ таблицы F размером $v \times v$, если $i \in S^{(3)}$, $j \in T^{(3)}$ и $i \neq j$, вписывается узел-граф $\begin{pmatrix} 4^{(6)} \\ [3;4;5^{(6)}] \end{pmatrix} 6^{(6)}_{[3;4^{(6)}]}$ (центр).

Шаг 2. В ячейки m_i таблицы-столбца M размером $v \times 1$, если $i \in T^{(1)}$, вписывается узел-граф $\begin{pmatrix} 4^{(1)} \\ [2;4^{(1)}] \end{pmatrix} 1^{(1)}_{[2;4;7^{(1)}]}$ (периферия).

Шаг 3. В ячейки l_j таблицы-строки L размером $1 \times v$, если $j \in S^{(7)}$, вписывается узел-граф $\begin{pmatrix} 4^{(5)} \\ [3;7^{(5)}] \end{pmatrix} 5^{(5)}_{[3;6;7^{(5)}]}$ (периферия).

Шаг 4. Если ячейка $f_{i,j}$ на пересечении строки i со столбцом j таблицы F не пуста, то выполняем регламент операции добавления узлов-графов.

В результате выполнения шагов 1–3 имеем.

		0	0	$\begin{pmatrix} 4^{(5)} \\ [3;7^{(5)}] \end{pmatrix} 5^{(5)}_{[3;6;7^{(5)}]}$	0	0	0	$\begin{pmatrix} 4^{(5)} \\ [3;7^{(5)}] \end{pmatrix} 5^{(5)}_{[3;6;7^{(5)}]}$
		1	2	3	4	5	6	7
	1	0	0	0	0	0	0	0
$\begin{pmatrix} 4^{(1)} \\ [2;4^{(1)}] \end{pmatrix} 1^{(1)}_{[2;4;7^{(1)}]}$	2	0	0	0	0	0	0	0
	3	0	0	0	$\begin{pmatrix} 4^{(6)} \\ [3;4;5^{(6)}] \end{pmatrix} 6^{(6)}_{[3;4^{(6)}]}$	0	0	0
$\begin{pmatrix} 4^{(1)} \\ [2;4^{(1)}] \end{pmatrix} 1^{(1)}_{[2;4;7^{(1)}]}$	4	0	0	$\begin{pmatrix} 4^{(6)} \\ [3;4;5^{(6)}] \end{pmatrix} 6^{(6)}_{[3;4^{(6)}]}$	0	0	0	0
	5	0		$\begin{pmatrix} 4^{(6)} \\ [3;4;5^{(6)}] \end{pmatrix} 6^{(6)}_{[3;4^{(6)}]}$	$\begin{pmatrix} 4^{(6)} \\ [3;4;5^{(6)}] \end{pmatrix} 6^{(6)}_{[3;4^{(6)}]}$	0		0
	6	0		0	0	0	0	0
$\begin{pmatrix} 4^{(1)} \\ [2;4^{(1)}] \end{pmatrix} 1^{(1)}_{[2;4;7^{(1)}]}$	7	0	0	0	0	0	0	0

Действия на шаге 4 продемонстрируем для ячейки $f_{4,3}$, так как только эта ячейка удовлетворяет условиям выполнения этого шага.

1. $d = \max w + 1 = 7 + 1 = 8$;
2. $U^{(8)} = U^{(1)} \parallel (i + U^{(6)} + j) \parallel U^{(5)} = 1 \parallel (4 + 6 + 3) \parallel 5 = 1,4,6,3,5$;
3. $S^{(8)} = S^{(1)} \setminus (S^{(1)} \cap U^{(8)}) = 2;4 \setminus (2;4 \cap 1;4;6;3;5) = 2;4 \setminus 4 = 2$;
4. $T^{(8)} = T^{(5)} \setminus (T^{(5)} \cap U^{(8)}) = 3;6;7 \setminus (3;6;7 \cap 1;4;6;3;5) = 3;6;7 \setminus 6;3 = 7;$

5. $\begin{pmatrix}1^{(8)}\\ 2^{(8)}\end{pmatrix}1,4,6,3,5\begin{bmatrix}8)\\ 7^{(8)}\end{bmatrix}$ существует;

6. $q^{(8)}=1$.

3.2.3. Пересечение узлов-графов. Для пояснения операции пересечения узлов-графов элементы подмножеств S и T будут располагаться вертикально без разделительных знаков.

Тогда граф на рисунке 1 будет представляться совокупностью узлов-графов следующего вида.

$$\begin{bmatrix}(4)\\2\\4\end{bmatrix}1\begin{bmatrix}(1)\\2\\4\\7\end{bmatrix}\quad\begin{bmatrix}(6)\\1\\3\\7\end{bmatrix}2\begin{bmatrix}(2)\\1\\3\\7\end{bmatrix}\quad\begin{bmatrix}(9)\\2\\5\\6\end{bmatrix}3\begin{bmatrix}(3)\\2\\4\\5\\6\end{bmatrix}\quad\begin{bmatrix}(4)\\1\\3\\6\end{bmatrix}4\begin{bmatrix}(4)\\1\\6\end{bmatrix}\quad\begin{bmatrix}(4)\\3\\7\end{bmatrix}5\begin{bmatrix}(5)\\3\\6\\7\end{bmatrix}\quad\begin{bmatrix}(4)\\3\\4\\5\end{bmatrix}6\begin{bmatrix}(6)\\3\\4\end{bmatrix}\quad\begin{bmatrix}(4)\\1\\2\\5\end{bmatrix}7\begin{bmatrix}(7)\\2\\5\end{bmatrix}$$

Предварительные пояснения. Пусть дан узел-граф ${}^{(q)}_{[S]}U^{(w)}_{[T]}$, у которого $u=2$. Разобьем тело узла-графа на два сегмента $U^{(w)}=\|_{i=1}^{2}\underset{\{i\}}{U}$. Между сегментами вставим подмножество несвязанных вершин $\underset{\{j\}}{R}$, через которые существуют пути длиной $p=2$ между $\underset{\{1\}}{U}$ и $\underset{\{2\}}{U}$. Такой узел-граф имеет вид, представленный на рисунке 3.

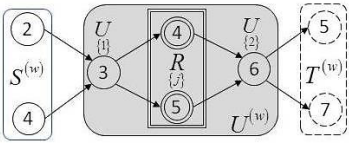


Рис. 3. Графическое изображение узла-графа

Условия существования такого узла-графа несколько расширяются:

– при $S^{(w)}\neq\varnothing\wedge T^{(w)}\neq\varnothing\wedge\underset{\{j\}}{R}\neq\varnothing$;

– при $s=t=1$ и $S^{(w)}\neq T^{(w)}$, и при $s=r=1$ и $S^{(w)}\neq\underset{\{j\}}{R}$, и при $t=r=1$ $T^{(w)}\neq\underset{\{j\}}{R}$, где r – мощность множества $\underset{\{j\}}{R}$.

Пусть имеются три узла-графа ${}^{(q)}_{[S]}U^{(1)}_{[T]}$, ${}^{(q)}_{[S]}U^{(2)}_{[T]}$ и ${}^{(q)}_{[S]}U^{(3)}_{[T]}$. Пересечением узлов-графов называется узел-граф ${}^{(q)}_{[S]}U^{(w=d)}_{[T]}={}^{(q)}_{[S]}U^{(1)}_{[T]}\cap{}^{(q)}_{[S]}U^{(2)}_{[T]}\cap{}^{(q)}_{[S]}U^{(3)}_{[T]}$, полученный по регламенту пересечения, представленному в таблице 4.

Таблица 4. Регламент пересечения узлов-графов

№ п/п	Регламент		Операции
1.	$d = \max w + 1$		$+$ – сложение
2.	$\underset{\{1\}}{R} = \left\langle T^{(1)} \cap S^{(2)} \right\rangle \setminus U^{(3)}$	Если $u^{(1)} \geq 2$, то вместо $U^{(1)}$ берется $\Big\Vert_{i=1}^{i=u^{(1)}} \underset{\{i\}}{U}$	$\setminus$ – разность множеств $\cap$ – пересечение множеств $\Vert$ – добавления сегмента тела узла-графа
3.	$\underset{\{2\}}{R} = \left\langle T^{(2)} \cap S^{(3)} \right\rangle \setminus U^{(1)}$		
4.	$S^{(w=d)} = \left[\left\{ S^{(1)} \setminus (S^{(1)} \cap U^{(2)}) \right\} \setminus \setminus (S^{(1)} \cap U^{(3)}) \right]$	Если $u^{(2)} \geq 2$, то вместо $U^{(2)}$ берется $\Big\Vert_{i=1}^{i=u^{(2)}} \underset{\{i\}}{U}$	
5.	$T^{(w=d)} = \left[\left\{ T^{(3)} \setminus (T^{(3)} \cap U^{(1)}) \right\} \setminus \setminus (T^{(3)} \cap U^{(2)}) \right]$	Если $u^{(3)} \geq 2$, то вместо $U^{(3)}$ берется $\Big\Vert_{i=1}^{i=u^{(3)}} \underset{\{i\}}{U}$	
6.	Если $s^{(w=d)} = 1$, то $\underset{\{1\}}{R} = \underset{\{1\}}{R} \setminus S^{(w=d)}$, $\underset{\{2\}}{R} = \underset{\{2\}}{R} \setminus S^{(w=d)}$, $T^{(w=d)} = T^{(w=d)} \setminus S^{(w=d)}$		
7.	Если $t^{(w=d)} = 1$, то $\underset{\{1\}}{R} = \underset{\{1\}}{R} \setminus T^{(w=d)}$, $\underset{\{2\}}{R} = \underset{\{2\}}{R} \setminus T^{(w=d)}$, $S^{(w=d)} = S^{(w=d)} \setminus T^{(w=d)}$		
8.	Если $\underset{\{1\}}{r} = 1$, то $\underset{\{2\}}{R} = \underset{\{2\}}{R} \setminus \underset{\{1\}}{R}$, $S^{(w=d)} = S^{(w=d)} \setminus \underset{\{1\}}{R}$, $T^{(w=d)} = T^{(w=d)} \setminus \underset{\{1\}}{R}$		
9.	Если $\underset{\{2\}}{r} = 1$, то $\underset{\{1\}}{R} = \underset{\{1\}}{R} \setminus \underset{\{2\}}{R}$, $S^{(w=d)} = S^{(w=d)} \setminus \underset{\{2\}}{R}$, $T^{(w=d)} = T^{(w=d)} \setminus \underset{\{2\}}{R}$		
10.	$U^{(w=d)} = \underset{\{1\}}{U}, \underset{\{1\}}{\langle R \rangle}, \underset{\{2\}}{U}, \underset{\{2\}}{\langle R \rangle}, \underset{\{3\}}{U}$		, – признак смежности
11.	Проверка условий существования узла-графа $\underset{[s]}{\overset{(q)}{U}}\underset{[T]}{\overset{(w=d)}{}}$		удаление узла-графа, если условие не выполняется
12.	Пересчет значения параметра q узла-графа		

Ограничения на операцию такие же, что и для операции добавления узлов-графов.

Пересечение двух узлов-графов с любой из сторон является частным случаем пересечения узлов-графов.

Выполним операцию пересечения, например для узлов-графов под номерами 1, 6 и 5 (рисунок 1), и получим:

1. $d = \max w + 1 = 7 + 1 = 8$;

2. $\underset{\{1\}}{R} = \left\langle T^{(1)} \cap S^{(6)} \right\rangle \setminus U^{(5)} = \left\langle \begin{bmatrix} 2 \\ 4 \\ 7 \end{bmatrix} \cap \begin{bmatrix} 3 \\ 4 \\ 6 \end{bmatrix} \right\rangle \setminus 5 = \langle 4 \rangle \setminus 5 = \underset{\{1\}}{\langle 4 \rangle}$;

3. $\underset{\{2\}}{R} = \left\langle T^{(6)} \cap S^{(5)} \right\rangle \setminus U^{(1)} = \left\langle \begin{bmatrix} 3 \\ 4 \end{bmatrix} \cap \begin{bmatrix} 3 \\ 7 \end{bmatrix} \right\rangle \setminus 1 = \langle 3 \rangle \setminus 1 = \underset{\{2\}}{\langle 3 \rangle}$;

4. $S^{(8)} = \left[\left\{ S^{(1)} \setminus (S^{(1)} \cap U^{(6)}) \right\} \setminus (S^{(1)} \cap U^{(5)}) \right] = \left[\left\{ \begin{bmatrix} 2 \\ 4 \end{bmatrix} \setminus \left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} \cap 6 \right) \right\} \setminus \right.$

$\left. \setminus \left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} \cap 5 \right) \right] = \left[\left\{ \begin{bmatrix} 2 \\ 4 \end{bmatrix} \setminus \varnothing \right\} \setminus \varnothing \right] = \begin{bmatrix} 2 \\ 4 \end{bmatrix}^{(8)}$;

5. $T^{(8)} = \left[\left\{ T^{(5)} \setminus (T^{(5)} \cap U^{(1)}) \right\} \setminus (T^{(5)} \cap U^{(6)}) \right] =$

$$= \left[\left\{ \begin{bmatrix} 2 \\ 3 \\ 6 \\ 7 \end{bmatrix} \setminus \left(\begin{bmatrix} 2 \\ 3 \\ 6 \\ 7 \end{bmatrix} \cap 1 \right) \right\} \setminus \left(\begin{bmatrix} 2 \\ 3 \\ 6 \\ 7 \end{bmatrix} \cap 6 \right) \right] = \left[\left\{ \begin{bmatrix} 2 \\ 3 \\ 6 \\ 7 \end{bmatrix} \setminus \varnothing \right\} \setminus 6 \right] = \begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix}^{(8)};$$

8. $\underset{(1)}{r} = 1$, $\underset{\{2\}}{R} = \underset{\{2\}}{R} \setminus \underset{\{1\}}{R} = \underset{\{2\}}{\langle 3 \rangle} \setminus \underset{\{1\}}{\langle 4 \rangle} = \underset{\{2\}}{\langle 3 \rangle}$, $S^{(8)} = S^{(8)} \setminus \underset{\{1\}}{R} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \setminus \langle 4 \rangle = [2]^{(8)}$,

$$T^{(8)} = T^{(8)} \setminus \underset{\{1\}}{R} = \begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix} \setminus \langle 4 \rangle = \begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix}^{(8)};$$

6. $s^{(8)} = 1$, $\underset{\{1\}}{R} = \underset{\{1\}}{R} \setminus S^{(8)} = \langle 4 \rangle \setminus [2] = \underset{\{1\}}{\langle 4 \rangle}$, $\underset{\{2\}}{R} = \underset{\{2\}}{R} \setminus S^{(8)} = \langle 3 \rangle \setminus [2] = \underset{\{2\}}{\langle 3 \rangle}$,

$$T^{(8)} = T^{(8)} \setminus S^{(8)} = \begin{bmatrix} 2 \\ 3 \\ 7 \end{bmatrix} \setminus [2] = \begin{bmatrix} 3 \\ 7 \end{bmatrix}^{(8)};$$

9. $\underset{\{2\}}{r}=1$, $\underset{\{1\}}{R}=\underset{\{1\}}{R}\setminus\underset{\{2\}}{R}=\langle 4\rangle\setminus\langle 3\rangle=\underset{\{1\}}{\langle 4\rangle}$, $S^{(8)}=S^{(8)}\setminus\underset{\{2\}}{R}=[2]\setminus\langle 3\rangle=[2]^{(8)}$,

$T^{(8)}=T^{(8)}\setminus\underset{\{2\}}{R}=\begin{bmatrix}3\\7\end{bmatrix}\setminus\langle 3\rangle=[7]^{(8)}$;

7. $t^{(8)}=1$, $\underset{\{1\}}{R}=\underset{\{1\}}{R}\setminus T^{(8)}=\langle 4\rangle\setminus[7]=\underset{\{1\}}{\langle 4\rangle}$, $\underset{\{2\}}{R}=\underset{\{2\}}{R}\setminus T^{(8)}=\langle 3\rangle\setminus[7]=\underset{\{2\}}{\langle 3\rangle}$,

$S^{(8)}=S^{(8)}\setminus T^{(8)}=[2]\setminus[7]=[2]^{(8)}$;

10. $U^{(8)}=\underset{\{1\}}{U}_{,\langle R\rangle_{\{1\}}}\underset{\{6\}}{U}_{,\langle R\rangle_{\{2\}}}\underset{\{5\}}{U}=\underset{\{1\}}{1},\underset{\{1\}}{\langle 4\rangle},\underset{\{6\}}{6},\underset{\{2\}}{\langle 3\rangle},\underset{\{5\}}{5}$;

11. $\underset{[s]}{{}^{(q)}}U^{(8)}_{[T]}=\underset{[2]}{{}^{(1)}}1,\langle 4\rangle,6,\langle 3\rangle,5^{(8)}_{[7]}$ – существует.

12. $q^{(8)}=1$.

3.3. Предварительная оценка метода трансформации узлов-графов. Здесь представлены соотношения, проверенные эмпирическим методом при решении очень большого числа задач на графах. Для получения верхней оценки сложности решения задачи перечисления гамильтоновых путей рассмотрен полный ориентированный граф $G[V,E]$. Для пересчета всех гамильтоновых путей в графе использовалась операция пересечения узлов-графов и получено следующее.

Граф задается v узлами-графами, которые объединяют пути длиной $p=2$, то есть количество узлов-графов (пространство перебора) на входе задачи составит $b_{(p=2)}=v$. Выполнение операции пересечения позволяет перейти от путей длиной $p=2$ к путям длиной $p=6$, потом к путям длиной $p=10$ и т.д. То есть минимальный размер шага $k=4$, его и будем рассматривать. Для поиска всех путей в графе с каждым из узлов-графов производится операция пересечения с каждой парой узлов-графов, у которых $U^{(1)}\cap U^{(2)}=\varnothing$, $U^{(1)}\cap U^{(3)}=\varnothing$, $U^{(2)}\cap U^{(3)}=\varnothing$. При поиске путей длиной $p=p+k$ количество таких пар составит $\left(v-\frac{p-4}{2}\right)^2-v+\frac{p-4}{2}$. Из этого следует, что число узлов-графов, объединяющих пути длиной

$p = p + k$, полученных при выполнении операции пересечения, будет составлять $b_{(p)} = b_{(p-4)}\left[\left(v - \frac{p-4}{2}\right)^2 - v + \frac{p-4}{2}\right]$.

Тогда для графов с мощностью множества вершин $v = 7,11,15,19,...$ и $k = 4$ имеем:

$$b_{(p)} = \begin{cases} v \text{ при } p = 2; \\ b_{(p-4)}\left[\left(v - \frac{p-4}{2}\right)^2 - v + \frac{p-4}{2}\right] \text{ при } p = 6,10,14,18,... . \end{cases}$$

Ограничение, накладываемое на v, призвано упростить расчетные выражения, но принципиальные зависимости для графов с произвольными мощностями вершин v сохраняются.

Общее количество объектов перебора при решении задачи перечисления гамильтоновых путей составит $B(v) = \sum_{p=2,6,10...(v-1)} b_{(p)}$, то есть менее чем $0,019\frac{v!}{\left(\frac{(v+1)}{2}\right)!}$. Очевидно, что пространство перебора сократилось чуть более чем в $\left(\frac{(v+1)}{2}\right)!$ раз. Например, для полного ориентированного графа, у которого мощность множества вершин составляет $v = 39$, количество комбинаторных объектов, участвующих в переборе путей, уменьшается более чем в $1,2 \cdot 10^{17}$ раз. А это очень значительное сокращение.

Нетрудно определить, что каждый узел-граф, например с мощностью $v = 7$, при представлении полного графа $G[V,E]$ объединяет $a_{(p=2)} = 30$ путей, а узел-граф, полученный после выполнения операции пересечения, объединяет $a_{(p=6)} = 24$ гамильтонова пути. Эти результаты связываются выражением $a_{(p)} = \frac{\left(v - \frac{p}{2}\right)!}{(v-p-1)!}$, справедливым для полных графов мощности $v = 7,11,15,19,...$ и путей длины $p = 6,10,14,18,...$

Общее количество гамильтоновых путей в полном ориентированном графе составляет $g = v!$, а для графа, состоящего из 7 вершин, $g = 5040$.

Для этого же графа при поиске путей методом трансформации получается $b_{(p=6)} = 210$ узлов-графов, каждый из которых объединяет $a_{(p=6)} = 24$ гамильтонова пути. Тогда $g = b_{(p=6)} a_{(p=6)} = 5040$. Таким образом, метод трансформации узлов-графов позволяет точно решать задачу перечисления гамильтоновых путей в графе.

4. Заключение. Принципиальными особенностями предлагаемого метода трансформации узлов-графов являются следующие факторы:

1. В отличие от эвристик, которые основываются на предположениях о наиболее перспективных ветвях процесса перебора:

– для сокращения пространства перебора используется разбиение множеств V и E графа $G[V, E]$ на более крупные комбинаторные объекты – узлы-графы;

– производится точное решение задачи.

2. В отличие от методов BFS и DFS, перебор производится путем центрально-периферического поглощения комбинаторных объектов: смежных узлов-графов, вершин и дуг.

По предварительной оценке, метод трансформации узлов-графов с размером шага $k = 4$ позволяет значительно ускорить перебор гамильтоновых путей при сохранении точности решения задачи.

Как и любой метод перебора, метод трансформации узлов-графов обладает свойством универсальности. То есть метод трансформации узлов-графов можно использовать для разработки множества алгоритмов:

– нахождения хотя бы одного решения задачи (задача существования);

– получения всех решений задачи;

– поиска оптимального решения задачи;

– подсчета количества решений задачи.

Однако для отдельных задач анализа сложных систем, например задачи расчета структурной надежности по совокупности путей или сечений [21], потребуется некоторое развитие правил преобразования выражений символьного умножения для учета пересечений путей по общим элементам.

Литература

1. Monch C., Rizk A. Directed Acyclic Graph-Type Distributed Ledgers via Young-Age Preferential Attachment // Stochastic Systems. 2023. vol. 13. no. 3. pp. 377–397.
2. Chopra S., Park H., Shim S. Extended Graph Formulation for the Inequity Aversion Pricing Problem on Social Networks // INFORMS Journal on Computing. 2022. vol. 34. no. 3. pp. 1327–1344.
3. Vidal T., Martinelli R., Pham T.A., Ha M.H. Arc Routing with Time-Dependent Travel Times and Paths // Transportation 2021 Science. vol. 55. no. 3. pp. 706–724.
4. Гасников А.В. Об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях // Математическое моделирование: 2015. Т. 27. № 12. С. 121–136.
5. Волков А.С., Баскаков А.Е. Разработка алгоритма многопутевой маршрутизации в программно-конфигурируемых сетях связи // T-Comm: Телекоммуникации и транспорт. 2021. Т. 15. № 9. С. 17–23.
6. Ray A., Ventresca M., Kannan K. A Graph-Based Ant Algorithm for the Winner Determination Problem in Combinatorial Auctions // Information Systems Research. 2021. vol. 32. no. 4. pp. 1099–1114.
7. Glasserman P., de Larrea E.L. Maximum Entropy Distributions with Applications to Graph Simulation. Operations Research. 2023. vol. 71. no. 5. pp. 1908–1924.
8. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи // М. Мир. 1982. 419 с.
9. Батенков К.А. Точные и граничные оценки вероятностей связности сетей связи на основе метода полного перебора типовых состояний // Труды СПИИРАН. 2019. Т. 18. № 5. С. 1093–1118.
10. Тимошенко А.В., Кочкаров Р.А., Кочкаров А.А. Выделение условий разрешимости NP-полных задач для класса предфрактальных графов // Моделирование и анализ информационных систем. 2021. Т. 28. № 2. С. 126–135.
11. Левин Л.А. Универсальные задачи перебора // Проблемы передачи информации. 1973. Т. 9. № 3. С. 115–116.
12. Gavanelli M., Mancini T. RCRA 2007: Experimental evaluation of algorithms for solving problems with combinatorial explosion. Journal of Algorithms. 2008. vol. 63. no. 1–3. pp. 1–2. DOI: 10.1016/j.jalgor.2008.02.002.
13. Кристофидес Н. Теория графов Алгоритмический подход // М.: Мир, 1978. 432 с.
14. Kudelia V.N. Full enumeration methods on graphs // T-Comm. 2023. vol. 17. № 7. pp. 57–64. DOI: 10.36724/2072-8735-2023-17-7-57-64.
15. Куделя В.Н. Методы перечисления путей в графе // Наукоемкие технологии в космических исследованиях Земли. 2023. Т. 15. № 5. С. 28–38.
16. Ott F., Markovic D., Strobel A., Kiebel S.J. Dynamic integration of forward planning and heuristic preferences during multiple goal pursuit // PLOS Computational Biology. 2020. vol. 16. no. 2. DOI: 10.1371/journal.pcbi.1007685.
17. Banville F., Gravel D., Poisot T. What constrains food webs? A maximum entropy framework for predicting their structure with minimal biases // PLOS Computational Biology. 2023. vol. 19. no. 9. DOI: 10.1371/journal.pcbi.1011458.
18. Harary F., Palmer E.M. Graphical enumeration. Academic Press New York and London, 1972. 271 p.
19. Kelil A, Dubreuil B, Levy E.D, Michnick S.W. Exhaustive search of linear information encoding protein-peptide recognition. PLOS Computational Biology. 2017. vol. 13. no. 4. DOI: 10.1371/journal.pcbi.1005499.
20. Рыбалов А.Н. О генерической сложности проблемы распознавания гамильтоновых путей // Прикладная дискретная математика. 2021. № 53. С. 120–126.

21. Мизин И.А., Богатырев В.А., Кулешов А.П. Сети коммутации пакетов // М. Радио и связь. 1986. 408 с.

Куделя Виктор Николаевич — ведущий специалист, АО «Институт Сетевых Технологий»; профессор кафедры, кафедра интеллектуальных систем автоматизации и управления, Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича» (СПбГУТ). Область научных интересов: сети передачи данных, применение методов теории графов и информатики в системах маршрутизации и распределения потоков, устойчивость (надежность, живучесть) систем, алгоритмическое обеспечение автоматизированных систем. Число научных публикаций — 106. Kudelia.Viktor@int.spb.ru; 17-я линия В.О., 54-1, 199178, Санкт-Петербург, Россия; р.т.: +7(812)331-8393,,215.

DOI 10.15622/ia.23.6.3

V. KUDELIA

SOLVING PATHS SEARCH PROBLEMS IN COMPLEX GRAPHS

Kudelia V. **Solving Paths Search Problems in Complex Graphs.**

Abstract. The construction of models of various systems is associated with the enumeration of the values of the parameters of the elements of the structure and taking into account all the characteristics of operation and interaction of components to find a certain set of solutions that determine the configuration of the system. Such tasks belong to enumeration type tasks and imply that some of the next solutions from this set are obtained from the previous solution in a certain order. It is known that any problem of the enumeration type is solved only by methods of exhaustive search, and other methods for their enumeration do not exist yet. The paper presents a new method of searching paths in a graph – the method of node-graph transformation. The proposed method, unlike the existing ones, allows one to search all directed simple paths in an oriented graph of arbitrary structure much faster. In the known graph search methods (Breadth First Search and Depth First Search), the object of the search is a path. The total number of such paths in the graph determines the size of the search space. The main idea of the node-graph transformation method is to significantly reduce the size of the search space by enlarging the search objects. The enlargement of enumeration objects is performed by clustering paths into combinatorial objects that concentrate some set of paths of the same length according to certain rules. These combinatorial objects are called node-graphs. A node-graph refers to center-peripheral combinatorial objects, and specific node-graph transformation operations have been developed to enumerate all paths in the graph, which allow finding new paths based on previous paths. The method can be used as a basic toolkit to reduce the dimensionality of the search space for solutions to NP-complete problems while maintaining the universality and accuracy of the full search.

Keywords: graph, enumeration, path, Hamiltonian path, combinatorial explosion, exhaustive search, NP-complete problem, clustering, node-graph.

References

1. Monch C., Rizk A. Directed Acyclic Graph-Type Distributed Ledgers via Young-Age Preferential Attachment. Stochastic Systems. 2023. vol. 13. no. 3. pp. 377–397.
2. Chopra S., Park H., Shim S. Extended Graph Formulation for the Inequity Aversion Pricing Problem on Social Networks. INFORMS Journal on Computing. 2022. vol. 34. no. 3. pp. 1327–1344.
3. Vidal T., Martinelli R., Pham T.A., Ha M.H. Arc Routing with Time-Dependent Travel Times and Paths. Transportation 2021 Science. vol. 55. no. 3. pp. 706–724.
4. Gasnikov A.V. [On the effective computability of competitive equilibria in transport-economic models]. Matematicheskoe modelirovanie – Mathematical Modeling: 2015. vol. 27. no. 12. pp. 121–136. (In Russ.).
5. Volkov A.S., Baskakov A.E. [Development of multipath routing algorithm in software-configurable communication networks]. T-Comm: Telekommunikacii i transport – T-Comm: Telecommunications and Transportation. 2021. vol. 15. no. 9. pp. 17–23. (In Russ.).
6. Ray A., Ventresca M., Kannan K. A Graph-Based Ant Algorithm for the Winner Determination Problem in Combinatorial Auctions. Information Systems Research. 2021. vol. 32. no. 4. pp. 1099–1114.
7. Glasserman P., de Larrea E.L. Maximum Entropy Distributions with Applications to Graph Simulation. Operations Research. 2023. vol. 71. no. 5. pp. 1908–1924.

8. Gjeri M., Dzhonson D. Vychislitel'nye mashiny i trudnoreshaemye zadachi [Computing machines and intractable problems]. M.: Mir. 1982. 419 p. (In Russ.).
9. Batenkov K.A. [Accurate and boundary estimate of communication network connectivity probability based on model state complete enumeration method]. Trudy SPIIRAN – SPIIRAS Proceedings. 2019. vol. 18. no. 5. pp. 1093–1118. (In Russ.).
10. Timoshenko A.V., Kochkarov R.A., Kochkarov A.A. [Allocation of solvability conditions of NP-complete problems for a class of pre-fractal graphs]. Modelirovanie i analiz informacionnyh sistem sistem – Information systems modeling and analysis. 2021. vol. 28. no. 2. pp. 126–135. (In Russ.).
11. Levin L.A. [Universal problems of brute force]. Problemy peredachi informacii – Information transfer problems. 1973. vol. 9. no. 3. pp. 115–116. (In Russ.).
12. Gavanelli M., Mancini T. RCRA 2007: Experimental evaluation of algorithms for solving problems with combinatorial explosion. Journal of Algorithms. 2008. vol. 63. no. 1–3. pp. 1–2. DOI: 10.1016/j.jalgor.2008.02.002.
13. Kristofides N. Teorija grafov Algoritmicheskij podhod [Graph Theory Algorithmic Approach]. M.: Mir, 1978. 432 p. (In Russ.).
14. Kudelia V.N. Full enumeration methods on graphs. T-Comm. 2023. vol. 17. № 7. pp. 57–64. DOI: 10.36724/2072-8735-2023-17-7-57-64.
15. Kudelia V.N. [Methods for enumerating paths in a graph]. Naukoemkie tehnologii v kosmicheskih issledovanijah Zemli – H&ES Reserch. 2023. vol. 15. no. 5. pp. 28–38. (In Russ.).
16. Ott F., Markovic D., Strobel A., Kiebel S.J. Dynamic integration of forward planning and heuristic preferences during multiple goal pursuit. PLOS Computational Biology. 2020. vol. 16. no. 2. DOI: 10.1371/journal.pcbi.1007685.
17. Banville F., Gravel D., Poisot T. What constrains food webs? A maximum entropy framework for predicting their structure with minimal biases. PLOS Computational Biology. 2023. vol. 19. no. 9. DOI: 10.1371/journal.pcbi.1011458.
18. Harary F., Palmer E.M. Graphical enumeration. Academic Press New York and London, 1972. 271 p.
19. Kelil A, Dubreuil B, Levy E.D, Michnick S.W. Exhaustive search of linear information encoding protein-peptide recognition. PLOS Computational Biology. 2017. vol. 13. no. 4. DOI: 10.1371/journal.pcbi.1005499.
20. Rybalov A.N. [On the generic complexity of the Hamiltonian path recognition problem] Prikladnaja diskretnaja matematika – Applied Discrete Mathematics. 2021. no. 53. pp. 120–126. (In Russ.).
21. Mizin I.A., Bogatyrev V.A., Kuleshov A.P. Seti kommutacii paketov [Packet switching networks]. M.: Radio i svjaz', 1986. 408 p. (In Russ.).

Kudelia Victor — Leading specialist, «Institute of Network Technologies»; Professor of the department, Department of Intelligent Automation and Control Systems, Federal State Budget-Financed Educational Institution of Higher Education The Bonch-Bruevich Saint Petersburg State University of Telecommunications (SPbSUT). Research interests: data transmission networks, application of graph theory and computer science methods in routing and flow distribution systems, resilience computer network, algorithmic support of automated systems. The number of publications — 106. Kudelia.Viktor@int.spb.ru; 54-1, 17th line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)331-8393,,215.

УДК 532.5:517.2:519.254 **DOI** 10.15622/ia.23.6.4

Г.М. Водинчар, Л.К. Фещенко

ВЫЧИСЛИТЕЛЬНАЯ ТЕХНОЛОГИЯ ПОСТРОЕНИЯ КАСКАДНЫХ МОДЕЛЕЙ МАГНИТОГИДРОДИНАМИЧЕСКОЙ ТУРБУЛЕНТНОСТИ

Водинчар Г.М., Фещенко Л.К. **Вычислительная технология построения каскадных моделей магнитогидродинамической турбулентности.**

Аннотация. В работе рассматривается вычислительная технология построения одного вида моделей мелкомасштабной магнитогидродинамической турбулентности – каскадных моделей (shell models). Любая такая модель является системой обыкновенных квадратично-нелинейных дифференциальных уравнений с постоянными коэффициентами. Каждая фазовая переменная интерпретируется по абсолютной величине как мера интенсивности одного из полей турбулентной системы в определенном диапазоне пространственных масштабов (масштабной оболочке). Уравнения любой каскадной модели должны обладать несколькими квадратичными инвариантами, которые являются аналогами законов сохранения в идеальной магнитогидродинамике. Вывод уравнений модели заключается в получении таких выражений для постоянных коэффициентов, при которых наперед заданные квадратичные выражения действительно будут инвариантами. Вывод этих выражений вручную является достаточно громоздким и вероятность ошибок в формульных преобразованиях велика. Особенно это касается нелокальных моделей, в которых могут взаимодействовать далекие по величине масштабные оболочки. Новизна и оригинальность работы состоит в том, что авторами предложена вычислительная технология, которая позволяет автоматизировать процесс вывода уравнений каскадных моделей. Технология реализована с использованием методов компьютерной алгебры, что позволило получать параметрические классы моделей, в которых инвариантность заданных квадратичных форм выполняется абсолютно точно – в формульном виде. Определение значений параметров в полученном параметрическом классе моделей далее выполняется за счет согласования мер взаимодействия оболочек в модели с вероятностями их взаимодействия в реальной физической системе. Идея описанной технологии и ее реализация принадлежит авторам. Отдельные ее элементы публиковались авторами ранее, однако в настоящей работе впервые дается ее систематическое описание для моделей с комплексными фазовыми переменными и согласованием мер взаимодействия оболочек с вероятностями. Аналогичных работ других авторов ранее не было. Технология позволяет быстро и безошибочно генерировать уравнения новых нелокальных каскадных моделей турбулентности и может быть полезна специалистам, занимающимся моделированием турбулентных систем.

Ключевые слова: автоматизация моделирования, компьютерная алгебра, турбулентность, магнитогидродинамика, каскадные модели.

1. Введение. Турбулентное движение включает пространственно-временные структуры различных масштабов, оно хаотично как во времени, так и в пространстве [1, 2]. При описании турбулентных систем часто используют представление полей системы пространственными спектрами, т.е. система описывается в пространстве волновых векторов, или масштабов [2 – 5]. Это обладает рядом преимуществ по сравнению с описанием в физическом пространстве, поскольку позволяет разделить

три основных процесса, протекающих в системе: возбуждение течений внешним воздействием, нелинейное взаимодействие вихревых структур и диссипацию. Именно представление турбулентной системы в пространстве масштабов лежит в основе каскадных моделей (shell models) турбулентности, об автоматизированном построении которых идет речь в настоящей работе.

Прямое численное моделирование турбулентности требует использования огромного числа узлов пространственно-временной сетки и мощных вычислительных ресурсов. Если же интерес для исследователя представляют только некоторые свойства турбулентной системы, можно использовать упрощенные модели, к которым относятся и каскадные модели. Эти модели являются квадратично-нелинейными динамическими системами с постоянными коэффициентами относительно небольшой размерности и интенсивно используются для моделирования процессов переноса энергии от интегрального масштаба (размера турбулентной системы) до диссипативного масштаба и для моделирования распределения статистических характеристик турбулентного потока [2, 6 – 16].

В отсутствие линейных членов и внешних воздействий каждая каскадная модель магнитогидродинамической (МГД) турбулентности должна обладать некоторым набором квадратичных инвариантов. Эти инварианты являются аналогами законов сохранения, выполняющихся в идеальной МГД. Поэтому вывод каждой новой модели связан с получением таких выражений для коэффициентов квадратичных членов, при которых заданные квадратичные формы от фазовых переменных будут инвариантами, причем точными инвариантами. Это требует проведения громоздких, хотя и идейно простых, алгебраических преобразований. Такая громоздкость приводит к временным затратам и является потенциальным источником ошибок при получении выражений для коэффициентов.

Целью настоящей работы является построение вычислительной технологии, которая позволила бы автоматизировать процесс формульного расчета коэффициентов каскадных моделей, при которых заданные квадратичные выражения были бы точными инвариантами. Эти коэффициенты должны формироваться либо в формульном виде, как функции свободных параметров, либо в виде точных числовых выражений (не в виде десятичных приближений). Технология позволит в автоматизированнном режиме получать уравнения новых каскадных моделей и будет полезными вычислительным инструментом для специалистов, занимающихся изучением турбулентности с помощью

каскадных моделей. Под вычислительной технологией мы подразумеваем совокупность расчетных методик, алгоритмов и программ.

Поскольку речь идет о программировании алгебраических преобразований, ясно, что для реализации технологии необходимо использовать методы и системы компьютерной алгебры (computer algebra system – CAS). Идея такой технологии принадлежит авторам, отдельные ее элементы публиковались ранее [17, 18], однако систематическое описание дается впервые. Подчеркнем, что речь идет не о численном исследовании готовых каскадных моделей, а именно о составлении уравнений, получении их коэффициентов в формульном виде. Отметим также, что ранее было описано применение CAS для автоматизации процесса построения спектральных моделей геодинамо [19, 20]. Следует отметить, что в настоящее время применение CAS широко распространено в математике, естественных и технических науках [21 – 25]. Данная работа демонстрирует такую возможность их использования, как автоматизацию процесса составления математических моделей.

2. Общий вид каскадных моделей МГД-турбулентности. Наиболее естественными образом можно подойти к построению каскадных моделей, исходя из спектральной формы МГД-уравнений, которые фактически и дают описание турбулентной системы в пространстве масштабов.

Физическое поле F произвольного тензорного ранга можно представить в реальном физическом пространстве как функцию $F(\mathbf{x}, t)$ радиус-вектора $\mathbf{x}$ и времени t. В равносильной форме оно допускает представление в пространстве волновых векторов $\mathbf{k}$ (пространстве Фурье) с помощью функции $\hat{F}(\mathbf{k}, t)$, называемой пространственным спектром. Это соответствие дается парой преобразований Фурье [1, 2]:

$$\begin{aligned} \hat{F}(\mathbf{k}, t) &= \int_{\mathbb{R}^3} F(\mathbf{x}, t)\, \mathrm{e}^{-\imath \mathbf{k}\mathbf{x}}\, d\mathbf{x}, \\ F(\mathbf{x}, t) &= \frac{1}{8\pi^3} \int_{\mathbb{R}^3} \hat{F}(\mathbf{k}, t)\, \mathrm{e}^{\imath \mathbf{k}\mathbf{x}}\, d\mathbf{k}. \end{aligned} \tag{1}$$

МГД-уравнения для вязкой несжимаемой жидкости в физическом пространстве имеют вид [1]:

$$\begin{aligned}
&\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p + \nu \triangle \mathbf{v} + \frac{1}{\mu\mu_0\sigma}\mathrm{rot}\,\mathbf{B} \times \mathbf{B} + \mathbf{f},\\
&\frac{\partial \mathbf{B}}{\partial t} = \mathrm{rot}\ (\mathbf{v} \times \mathbf{B}) + \eta \triangle \mathbf{B},\\
&\nabla \mathbf{v} = 0, \quad \nabla \mathbf{B} = 0,
\end{aligned} \tag{2}$$

где поля $\mathbf{v}$, p, $\mathbf{B}$ и $\mathbf{f}$ – это скорость, давление, магнитная индукция и заданная массовая плотность внешних сил соответственно. Остальные параметры в системе постоянны.

Равносильные уравнения в пространстве волновых векторов для спектров полей имеют следующий вид [1, 2]:

$$\begin{aligned}
&\frac{\partial \hat{\mathbf{v}}}{\partial t} = \imath \int_{\mathbb{R}^3} d\mathbf{q} \int_{\mathbb{R}^3} d\mathbf{s}\, \delta(\mathbf{k}+\mathbf{q}+\mathbf{s})\, S(\mathbf{k},\mathbf{q},\mathbf{s}) \bullet \hat{\mathbf{v}}^*(\mathbf{q}) \bullet \hat{\mathbf{v}}^*(\mathbf{s}) +\\
&+\frac{\imath}{\mu\mu_0\sigma} \int_{\mathbb{R}^3} d\mathbf{q} \int_{\mathbb{R}^3} d\mathbf{s}\, \delta(\mathbf{k}+\mathbf{q}+\mathbf{s})\, L(\mathbf{k},\mathbf{q},\mathbf{s}) \bullet \hat{\mathbf{B}}^*(\mathbf{q}) \bullet \hat{\mathbf{B}}^*(\mathbf{s}) -\\
&-\nu k^2 \hat{\mathbf{v}} + \mathbf{k} \times \left(\mathbf{k} \times \hat{\mathbf{f}}\right)/k^2,\\
&\frac{\partial \hat{\mathbf{B}}}{\partial t} = \imath \int_{\mathbb{R}^3} d\mathbf{q} \int_{\mathbb{R}^3} d\mathbf{s}\, \delta(\mathbf{k}+\mathbf{q}+\mathbf{s})\, W(\mathbf{k},\mathbf{q},\mathbf{s}) \bullet \hat{\mathbf{v}}^*(\mathbf{q}) \bullet \hat{\mathbf{B}}^*(\mathbf{s}) -\\
&-\eta k^2 \hat{\mathbf{B}},
\end{aligned} \tag{3}$$

где $S(\cdot,\cdot,\cdot), L(\cdot,\cdot,\cdot), W(\cdot,\cdot,\cdot)$ – некоторые вещественные тензорные функции 3-го ранга (их конкретный вид неважен для настоящей работы), $k = \|\mathbf{k}\|$ – волновое число, $\delta(\cdot)$ – трехмерная δ-функция, символ $\bullet$ означает тензорную свертку, а звездочка – комплексное сопряжение.

В интегральных членах идет интегрирование (непрерывное суммирование) по всем волновым векторам $\mathbf{q}$ и $\mathbf{s}$. Формально имеет место взаимодействие любых трех волн, но из-за наличия δ-функции реально взаимодействуют только те волны, из волновых векторов которых можно составить треугольник. Это очень важно для построения каскадных моделей далее.

На спектральной форме МГД-уравнений (3) и основывается построение каскадных моделей.

Для этого прежде всего проводится иерархическое разбиение пространства волновых векторов на масштабные оболочки по следующей схеме [2, 6]. В качестве основного пространственного масштаба используется размер физической турбулентной системы, принимаемый за единицу длины. Также фиксируется число $q > 1$. Далее вводятся иерархические диапазоны масштабов $D_n = \left(q^{-n-1}; q^{-n}\right]$, где $n \in \mathbb{Z}$, и разбиение пространства волновых векторов на соответствующие оболочки $P_n = \left\{\mathbf{k} \;\middle|\; q^n \leqslant \|\mathbf{k}\| < q^{n+1}\right\}$. Ясно, что объединением всех диапазонов масштабов будет $(0; +\infty)$. Исключено лишь нулевое волновое число, связанное с бесконечно большим масштабом, который не представляет физического интереса.

Затем вводятся комплексные переменные $U_n(t)$ и $B_n(t)$, называемые коллективными, модули которых интерпретируются как суммарные меры интенсивности волновых структур скорости и магнитного поля, волновые векторы которых лежат в оболочке P_n. Направления волновых векторов и конкретные значения волновых чисел могут быть какими угодно, важно лишь, чтобы эти числа лежали в пределах заданных диапазонов. Идея каскадных моделей в том, чтобы турбулентную систему упрощенно описать в терминах динамики этих коллективных переменных. Тогда для них необходимо составить динамические уравнения. Эти уравнения можно получать различными способами [1, 2, 6], но физически наиболее естественный – это взять прямой аналог спектральных уравнений (3), предварительно их обезразмерив.

Тогда возникают общие уравнения каскадной модели:

$$\begin{aligned}
&\frac{dU_n}{dt} = \imath \sum_{l,m=-\infty}^{+\infty} S_{nlm} U_l^* U_m^* + \imath \sum_{l,m=-\infty}^{+\infty} L_{nlm} B_l^* B_m^* - \\
&-\mathsf{Re}^{-1} k_n^2 U_n + f_n(t), \\
&\frac{dB_n}{dt} = \imath \sum_{l,m=-\infty}^{+\infty} W_{nlm} U_l^* B_m^* - \mathsf{Rm}^{-1} k_n^2 B_n,
\end{aligned} \tag{4}$$

где $k_n = q^n$ – характерное волновое число n-ой оболочки, S_{nlm}, L_{nlm}, W_{nlm} – некоторые постоянные вещественные коэффициенты (коэффициенты нелинейных взаимодействий), заданная функция $f_n(t)$ моделирует внешний подвод энергии в n-ую оболочку, Re – число

Рейнольдса, Rm – магнитное число Рейнольдса. Обычно полагают $f_n(t) \equiv 0$ при $n \neq 0$, что соответствует подводу энергии только на основном масштабе турбулентной системы.

Прямая аналогия между спектральными уравнениями (3) и каскадными уравнениями (4) очевидна. Такая форма каскадных уравнений известна как модели типа GOY [1, 2, 6]. Модели этого типа отличаются друг от друга коэффициентами нелинейных взаимодействий. Именно технология их расчета с помощью символьных вычислений, с учетом тех или иных ограничений, является предметом настоящей работы. Кроме моделей типа GOY, часто используют еще и модели типа Sabra [1, 2, 8]. Их отличие от (4) в том, что в каждом квадратичном члене либо только одна фазовая переменная входит в сопряжении, либо обе без сопряжения. Предлагаемая в данной работе технология расчета коэффициентов для моделей типа GOY легко может быть модифицирована для Sabra, а также для комбинированных моделей.

Коэффициенты квадратичных членов уравнений (4) определяются тремя индексами, первый из которых соответствует номеру уравнения, то есть номеру выделенной масштабной оболочки, а два других – номерам оболочек, взаимодействующих с выделенной, т.е. имеется некоторая неоднородность в пространстве масштабов. Однако в реальности для нелинейного переноса энергии в пределах инерционного интервала существует масштабная однородность (самоподобие). В таком случае взаимодействие i-ой и j-ой оболочек с n-ой должно зависеть не от самих i и j, а только от их отличия от n. Кроме этого, все нелинейные члены уравнений в физическом пространстве (2) содержат оператор набла. Аналогом этого оператора в n-ой масштабной оболочке выступает волновое число k_n.

Поэтому модели (4) часто записывают в другой форме:

$$\begin{aligned}
&\frac{dU_n}{dt} = \imath k_n \sum_{i,j=-\infty}^{+\infty} \hat{S}_{ij} U^*_{n+i} U^*_{n+j} + \imath k_n \sum_{i,j=-\infty}^{+\infty} \hat{L}_{ij} B^*_{n+i} B^*_{n+j} - \\
&-\mathsf{Re}^{-1} k_n^2 U_n + f_n(t), \\
&\frac{dB_n}{dt} = \imath k_n \sum_{i,j=-\infty}^{+\infty} \hat{W}_{nij} U^*_{n+i} B^*_{n+j} - \mathsf{Rm}^{-1} k_n^2 B_n,
\end{aligned} \tag{5}$$

где матрицы коэффициентов $\hat{S}_{ij}$ и $\hat{L}_{ij}$ симметричные. Ясно, что соответствие между коэффициентами в (4) и (5) будет таким:

$$S_{nlm} = k_n \hat{S}_{ij}, \quad L_{nlm} = k_n \hat{L}_{ij}, \quad W_{nlm} = k_n \hat{W}_{ij}, \\ i = l - n, \quad j = m - n. \tag{6}$$

Первый вариант записи каскадной модели оказывается удобнее для расчета коэффициентов. Второй же вариант соответствует представлению о пространственном самоподобии турбулентного потока.

Число q, определяющее иерархию волновых оболочек P_n, довольно произвольно, но обычно берут в практике построения каскадных моделей либо $q = 2$, либо $q = \left(1 + \sqrt{5}\right)/2$ – «золотое сечение». У обоих этих вариантов есть свои преимущества. Первый хорошо отражает представление о турбулентности как об иерархической структуре из вложенных вихрей, когда каждый трехмерный вихрь распадается на восемь подвихрей вдвое меньшего размера. Второй оптимален с точки зрения эффективности переноса энергии в системе [2].

Рассмотрим теперь, какие ограничения на коэффициенты каскадных моделей накладывает требование существования квадратичных инвариантов. Известно, что при $\nu = \eta = 0$ и $\mathbf{f} \equiv 0$ МГД-уравнения (2) обладают следующими инвариантами:

– полная энергия

$$E = \frac{1}{2} \int_{\mathbb{R}^3} \left(\mathbf{v}^2 (\mathbf{x}, t) + \mathbf{B}^2 (\mathbf{x}, t) \right) \, d\mathbf{x}; \tag{7}$$

– перекрестная спиральность

$$H_C = \int_{\mathbb{R}^3} \mathbf{v} (\mathbf{x}, t) \cdot \mathbf{B} (\mathbf{x}, t) \, d\mathbf{x}; \tag{8}$$

– квадрат потенциала магнитного поля (для двумерных течений)

$$A^2 = \int_{\mathbb{R}^3} \mathbf{A}^2 (\mathbf{x}, t) \, d\mathbf{x}, \tag{9}$$

где $\mathbf{A} (\mathbf{x}, t)$ – потенциал поля магнитной индукции, определяемый равенствами $\operatorname{rot} \mathbf{A} = \mathbf{B}$, $\nabla \mathbf{A} = 0$;

– магнитная спиральность (для трехмерных течений)

$$H_B = \int_{\mathbb{R}^3} \mathbf{A}\left(\mathbf{x},t\right) \cdot \mathbf{B}\left(\mathbf{x},t\right)\, d\mathbf{x}. \tag{10}$$

Выражения для этих инвариантов через спектры полей имеют вид:

$$\begin{aligned}
E &= \frac{1}{16\pi^3} \int_{\mathbb{R}^3} \left(\hat{\mathbf{v}}\left(\mathbf{k},t\right) \cdot \hat{\mathbf{v}}^*\left(\mathbf{k},t\right) + \hat{\mathbf{B}}\left(\mathbf{k},t\right) \cdot \hat{\mathbf{B}}^*\left(\mathbf{k},t\right) \right) d\mathbf{k}, \\
H_C &= \frac{1}{8\pi^3} \int_{\mathbb{R}^3} \hat{\mathbf{v}}(\mathbf{k},t) \cdot \hat{\mathbf{B}}^*(\mathbf{k},t)\, d\mathbf{k}, \\
A^2 &= \frac{1}{8\pi^3} \int_{\mathbb{R}^3} \frac{1}{k^4} \left(\mathbf{k} \times \hat{\mathbf{B}}(\mathbf{k},t) \right) \cdot \left(\mathbf{k} \times \hat{\mathbf{B}}^*(\mathbf{k},t) \right) d\mathbf{k}, \\
H_B &= \frac{1}{8\pi^3} \int_{\mathbb{R}^3} \frac{\imath}{k^2} \cdot \left(\mathbf{k} \times \hat{\mathbf{B}}(\mathbf{k},t) \right) \cdot \hat{\mathbf{B}}^*(\mathbf{k},t)\, d\mathbf{k}.
\end{aligned} \tag{11}$$

В таком случае каскадные модели (4) и (5) тоже должны обладать при $\mathsf{Re}^{-1} = \mathsf{Rm}^{-1} = 0$ и $f_n(t) \equiv 0$ квадратичными по коллективным переменным инвариантами, являющимися каскадными аналогами (11). Для полной энергии, перекрестной спиральности и квадрата потенциала прямые аналоги очевидны, они определяются формулами (сохраняем обозначения инвариантов):

$$\begin{aligned}
E &= \frac{1}{2} \sum_{n=-\infty}^{\infty} \left(U_n U_n^* + B_n B_n^* \right), \\
H_C &= \sum_{n=-\infty}^{\infty} \left(U_n B_n^* + U_n^* B_n \right), \\
A^2 &= \sum_{n=-\infty}^{\infty} k_n^{-2} B_n B_n^*.
\end{aligned} \tag{12}$$

С инвариантом магнитной спиральности ситуация сложнее. Прямым ее аналогом по четвертой формуле (11) будет $\sum\limits_{n=-\infty}^{\infty} k_n^{-1} B_n B_n^*$, однако это выражение знакоположительное, а спиральность может быть любого знака. Поэтому можно рассматривать различные варианты

введения каскадной формы магнитной спиральности с помощью знакопеременных выражений, составленных из волновых чисел k_n и коллективных переменных магнитного поля $B_n(t)$. Простейший вариант – это искусственное введение знакопеременности и определение каскадной магнитной спиральности формулой

$$H_B = \sum_n (-1)^n k_n^{-1} B_n B_n^*. \tag{13}$$

Иногда рассматривают знакопеременное выражение более общего вида

$$H_B^\lambda = \sum_n (-1)^n k_n^{-\lambda} B_n B_n^*, \tag{14}$$

с произвольным $\lambda > 0$, называя его обобщенной спиральностью [2]. Ясно, что оно имеет размерность спиральности только при $\lambda = 1$ и совпадает в этом случае с (13).

3. Вычислительная технология для составления каскадной модели. Для получения уравнений конкретной модели в форме (4) или (5) необходимо найти коэффициенты нелинейных членов. Их расчет основан на ограничениях, которым должны удовлетворять уравнения модели. Эти ограничения следующие:

1. Существование квадратичных инвариантов в отсутствие диссипации и внешних источников, т.е. при $\mathsf{Re}^{-1} = \mathsf{Rm}^{-1} = 0$ и $f_n(t) \equiv 0$.

2. Возможность нелинейного взаимодействия оболочек P_n, P_{n+i} и P_{n+j}.

3. Предельная, допускаемая в модели, дальность взаимодействия в пространстве масштабов.

4. Существование степенных стационарных решений (не обязательно).

5. Согласование величины коэффициентов с вероятностями взаимодействия оболочек.

От каскадных моделей в отсутствие диссипации и внешних источников обычно требуют еще сохранение фазового объема, однако несложно показать, что модели типа GOY сохраняют его при любых коэффициентах нелинейных членов.

Далее мы опишем вычислительную технологию расчета коэффициентов, удовлетворяющих этим ограничениям. Учет первых четырех ограничений реализован с помощью методов компьютерной алгебры и дает классы моделей, в которых эти ограничения будут выполняться точно – на формульном уровне. Для случая каскадных моделей с вещественными коллективными переменными подобная технология была описана в [17]. Схема расчета для комплексных каскадных моделей гидродинамики и конвекции была описана в [18]. Случай комплексных каскадных моделей магнитогидродинамики описан в настоящей работе впервые.

После получения выражений для коэффициентов выбор конкретных значений параметров в этих выражениях проводится путем согласования коэффициентов с вероятностями взаимодействия оболочек с помощью численной оптимизации.

Общая схема разработанной технологии приведена на рисунке 1. Сначала кратко опишем ее блоки, а потом перейдем к детальному описанию.

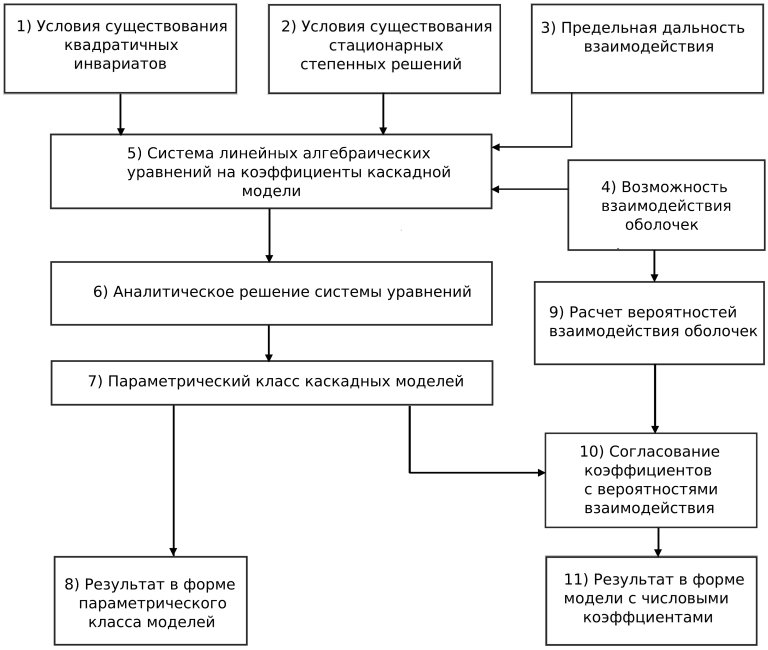


Рис. 1. Схема разработанной технологии составления каскадных моделей

Условия существования квадратичных инвариантов и стационарных степенных решений (блоки 1 и 2 на схеме) получаются в виде семейства линейных однородных алгебраических уравнений на коэффициенты нелинейных членов. Исходными данными здесь являются формулы инвариантов и стационарного решения, а на выходе получаются уравнения для коэффициентов.

Предельная дальность взаимодействия Q (блок 3 на схеме) является целочисленным задаваемым параметром.

Возможность взаимодействия оболочек P_n, P_{n+i}, P_{n+j} (блок 4 на схеме) задается в форме предиката $Q(i,j)$, определяющего их индексы и не зависящего от n. Формула предиката в технологии задана и определяется вручную по критерию: если в S_n, S_{n+i}, S_{n+j} существуют волновые векторы, из которых составляется треугольник, то оболочки могут взаимодействовать. Этот предикат определяет коэффициенты модели, которые могут быть отличны от нуля.

На основе выходов блоков 1-4 программно в CAS формируется система линейных алгебраических уравнений на коэффициенты (блок 5 на схеме). Затем средствами CAS ищется аналитическое решение данной системы (блок 6 на схеме). На выходе этого блока получается решение в виде выражений, зависящих от свободных параметров.

Выход блока 6 дает параметрический класс каскадных моделей, которые при любых значениях параметров обладают нужными инвариантами, стационарными решениями и предельной дальностью взаимодействий (блок 7 на схеме).

Если не ставится задача согласования коэффициентов с вероятностями, данный класс является итоговым результатом (блок 8 на схеме). В противном случае проводятся дальнейшие численные расчеты. Сначала методом Монте-Карло вычисляются вероятности взаимодействия оболочек P_n, P_{n+i}, P_{n+j}, т.е. вероятности того, что из трех векторов, случайно взятых по одному из каждой оболочки, можно составить треугольник (блок 9 на схеме).

По выходам блоков 7 и 9 численно выполняется минимизация относительных отклонений коэффициентов модели (мер взаимодействия оболочек) от вероятностей этих взаимодействий. Минимизация выполняется по параметрам (блок 10 на схеме). В результате получается модель с числовыми значениями коэффициентов (блок 11 на схеме).

Теперь опишем этапы технологии более подробно.

3.1. Условия существования квадратичных инвариантов. Дифференцирование квадратичных форм (12) и (13) по времени и подстановка вместо производных коллективных переменных их

выражений из правых частей (4) даст кубичные формы от коллективных переменных. Коэффициенты этих форм будут линейными комбинациями коэффициентов нелинейных взаимодействий. Поскольку тождественное равенство нулю кубичных форм возможно только при всех нулевых коэффициентах этих форм, будем получать набор однородных линейных уравнений для искомых коэффициентов нелинейных взаимодействий.

Практическая реализация этого требует весьма громоздких формульных преобразований, которые удобно выполнять, используя средства CAS, чтобы избежать ошибок в алгебраических преобразованиях. Мы будем использовать пакет Maple [26, 27].

Начнем с условия сохранения полной энергии. Для ее изменения получаем уравнение

$$\begin{aligned}\frac{dE}{dt} &= \frac{d}{dt}\frac{1}{2}\sum_n (U_n U_n^* + B_n B_n^*) = \\ &= \frac{1}{2}\sum_n \left(\frac{dU_n}{dt}U_n^* + \frac{dU_n^*}{dt}U_n + \frac{dB_n}{dt}B_n^* + \frac{dB_n^*}{dt}B_n\right).\end{aligned} \tag{15}$$

Далее рассмотрим выполнение необходимых преобразований в пакете Maple и результат работы пакета. Удобно ввести действительные и мнимые части коллективных переменных $U_n = x_n + \imath y_n$ и $B_n = g_n + \imath h_n$. Выполним в Maple следующий набор команд из Листинга 1.

```
assume(x[n], real, x[l], real, x[m], real):
assume(y[n], real, y[l], real, y[m], real):
assume(g[n], real, g[l], real, g[m], real):
assume(h[n], real, h[l], real, h[m], real):
assume(S[n, l, m], real, L[n, l, m], real, W[n, l, m], real):
u[n] := x[n] + I * y[n]: B[n] := g[n] + I * h[n]:
u[l] := x[l] + I * y[l]: B[l] := g[l] + I * h[l]:
u[m] := x[m] + I * y[m]: B[m] := g[m] + I * h[m]:
du[n] := I * S[n, l, m] * conjugate(u[l]) * conjugate(u[m]) +
I * L[n ,l ,m] * conjugate(B[l])  *conjugate(B[m]);
dB[n] := I *W[n,l,m] * conjugate(u[l]) * conjugate(B[m]);
```

Листинг 1. Определение выражений для коллективных переменных и уравнений модели

Команды вида assume(eq, real) в строках 1-5 указывают пакету, что переменная eq является вещественной. Это важно при

работе с операцией сопряжения, т.к. по умолчанию Maple считает все переменные комплексными. В строках 6-8 в соответствующих переменных формируются в общем виде выражения для коллективных переменных модели. Заметим, что мнимая единица i в Maple задается как системная константа I. Далее (строки 9-11) в переменных du[n] и dB[n] формируются правые части уравнений (4). Все символы суммирования здесь и далее опускаются, поскольку они излишне загромождали бы формируемые Maple выражения, а фактически они нигде не используются, поскольку работа идет только с одночленами форм.

Результат выполнения этих команд (вывод Maple) представлен на рисунке 2, который является распечаткой фрагмента экрана при сеансе работы с пакетом. Наличие символа тильда справа от имени переменной при выводе означает, что на переменную было наложено ранее ограничение с помощью команды assume.

$$du_n := \mathrm{I}\, S\sim_{n,l,m} \left(x\sim_l - \mathrm{I}\, y\sim_l\right)\left(x\sim_m - \mathrm{I}\, y\sim_m\right) + \mathrm{I}\, L\sim_{n,l,m} \left(g\sim_l - \mathrm{I}\, h\sim_l\right)\left(g\sim_m - \mathrm{I}\, h\sim_m\right)$$
$$dB_n := \mathrm{I}\, W\sim_{n,l,m} \left(x\sim_l - \mathrm{I}\, y\sim_l\right)\left(g\sim_m - \mathrm{I}\, h\sim_m\right)$$

Рис. 2. Результат выполнения Maple команд, определяющих алгебраические выражения для уравнений каскадной модели

Далее в переменной dE формируются члены кубичной формы условия сохранения энергии по формуле (15). Соответствующая команда Maple представлена в Листинге 2.

```
20 dE := 1/2 * simplify(
        du[n] * conjugate(u[n]) + u[n]  * conjugate(du[n]) +
22      dB[n] * conjugate(B[n]) + B[n]  * conjugate(dB[n])
        );
```

Листинг 2. Формирование кубичной формы для изменения энергии

Результат выполнения этой команды приведен на рисунке 3.

$$\begin{aligned} dE := {} & S\sim_{n,l,m} x\sim_l x\sim_m y\sim_n + S\sim_{n,l,m} x\sim_l y\sim_m x\sim_n + S\sim_{n,l,m} y\sim_l x\sim_m x\sim_n - S\sim_{n,l,m} y\sim_l y\sim_m y\sim_n \\ & + L\sim_{n,l,m} g\sim_l g\sim_m y\sim_n + L\sim_{n,l,m} g\sim_l h\sim_m x\sim_n + L\sim_{n,l,m} h\sim_l g\sim_m x\sim_n - L\sim_{n,l,m} h\sim_l h\sim_m y\sim_n \\ & + W\sim_{n,l,m} x\sim_l g\sim_m h\sim_n + W\sim_{n,l,m} x\sim_l h\sim_m g\sim_n + W\sim_{n,l,m} y\sim_l g\sim_m g\sim_n - W\sim_{n,l,m} y\sim_l h\sim_m h\sim_n \end{aligned}$$

Рис. 3. Результат построения Maple одночленов формы условия сохранения энергии

Из выражения на рисунке 3 видно, что после соответствующих перестановок индексов требуется обеспечить равенство нулю суммы пяти

кубичных форм: одной с одночленами $x_l x_m y_n$, второй с одночленами $y_l y_m y_n$, третей с одночленами $g_l g_m y_n$, четвертой с одночленами $h_l h_m y_n$ и пятой с одночленами $g_l h_m x_n$. Для этого необходимо и достаточно равенство нулю всех коэффициентов форм, однако надо учесть подобные слагаемые. Для первой формы подобные слагаемые получаются перестановками индексов l и m, поэтому получим условие

$$S_{nlm} + S_{nml} + S_{mln} + S_{lmn} + S_{lnm} + S_{mnl} = 0. \tag{16}$$

Ясно, что в ситуации, когда $l = m$, подобных слагаемых не будет и эти перестановки не нужны, однако их формальное выполнение просто приведет к умножению всех членов уравнения на два.

Для второй формы с одночленами $y_l y_m y_n$ подобные слагаемые получаются перестановками всех трех индексов. Получится то же самое уравнение, что и в предыдущем случае.

Для третьей формы с одночленами $g_l g_m y_n$ так же, как и для первого случая, подобные слагаемые получаются перестановками индексов l и m. Получится следующее условие

$$L_{nlm} + L_{nml} + W_{mnl} + W_{lnm} = 0. \tag{17}$$

Это же условие получается для двух оставшихся форм. Ясно, что и здесь при совпадении двух или всех трех индексов перестановки излишни, но их формальное выполнение просто приведет к умножению членов уравнения на два или шесть.

Теперь целесообразно перейти к двухиндексному варианту записи в соответствии с формулами (6). Получим из (16) и (17) уравнения на коэффициенты, обеспечивающие выполнение закона сохранения энергии

$$\begin{aligned} &q^n \hat{S}_{l-n,m-n} + q^n \hat{S}_{m-n,l-n} + q^l \hat{S}_{n-l,m-l} + q^l \hat{S}_{m-l,n-l} + \\ &+ q^m \hat{S}_{l-m,n-m} + q^m \hat{S}_{n-m,l-m} = 0 \\ &q^n \hat{L}_{l-n,m-n} + q^n \hat{L}_{m-n,l-n} + q^m \hat{W}_{n-m,l-m} + q^l \hat{W}_{n-l,m-l} = 0. \end{aligned} \tag{18}$$

Деля почленно на q^n и переходя к индексам (i, j) по схеме $i = l - n, j = m - n$, получаем

$$\begin{aligned}
&\hat{S}_{i,j} + \hat{S}_{j,i} + q^i \left(\hat{S}_{-i,j-i} + \hat{S}_{j-i,-i} \right) + \\
&+q^j \left(\hat{S}_{i-j,-j} + \hat{S}_{-j,i-j} \right) = 0, \\
&\hat{L}_{i,j} + \hat{L}_{j,i} + q^i \hat{W}_{-i,j-i} + q^j \hat{W}_{-j,i-j} = 0.
\end{aligned} \tag{19}$$

Итак, условие сохранения полной энергии дает два семейства уравнений (19).

Проводя похожие расчеты, можно получить уравнения, обеспечивающие и другие инварианты. Сами эти расчеты в тексте не описываем, приведем только итоговые уравнения.

Сохранение перекрестной спиральности H_C дает следующие уравнения:

$$\begin{aligned}
&\hat{L}_{i,j} + \hat{L}_{j,i} + q^i \left(\hat{L}_{-i,j-i} + \hat{L}_{j-i,-i} \right) + \\
&+q^j \left(\hat{L}_{i-j,-j} + \hat{L}_{-j,i-j} \right) = 0, \\
&\hat{S}_{i,j} + \hat{S}_{j,i} + q^i \hat{W}_{j-i,-i} + q^j \hat{W}_{i-j,-j} = 0.
\end{aligned} \tag{20}$$

Требование сохранения квадрата магнитного потенциала приводит к уравнениям:

$$\hat{W}_{i,i} + q^{-j} \hat{W}_{i-j,-j} = 0. \tag{21}$$

Наконец, условие сохранения магнитной спиральности H_B дает уравнения:

$$\hat{W}_{i,j} + (-1)^j \hat{W}_{i-j,-j} = 0. \tag{22}$$

Коллективные переменные каскадных моделей не содержат информации о геометрии движений, о том, моделируется ли двумерный или трехмерный поток. Однако двумерная и трехмерная турбулентность ведет себя существенно разным образом [1, 6, 28, 29]. Поэтому необходимо как-то различать эти случаи в каскадных моделях. Это можно сделать с помощью инвариантов. Поскольку в двумерном потоке

сохраняется квадрат потенциала поля, а в трехмерном – магнитная спиральность, то, потребовав от модели сохранения одного их этих выражений, можно косвенно внести в модель информацию о размерности движения.

Итак, на выходе блока 1 схемы с рисунке 1 получаем семейства уравнений (19, 20, 21, 22) на коэффициенты модели.

3.2. Условие существования стационарных степенных решений. Известно, что в однородной изотропной турбулентности в пределах итерационного интервала некоторые статистические характеристики обладают степенными законами распределения по масштабам, а значит, и по волновым числам [1]. Пусть $Q(k) \sim k^{\lambda}$ – одна из таких характеристики. Тогда связанное с коллективными переменными интегральное значение этой характеристики в n-ой оболочке будет

$$Q_n \sim \int\limits_{q^n}^{q^{n+1}} Q(k)\,dk \sim \int\limits_{q^n}^{q^{n+1}} k^{\lambda}\,dk \sim q^{n(\lambda+1)}. \tag{23}$$

В качестве примера наиболее известного спектрального закона возьмем закон Колмогорова для энергии $E(k) \sim k^{-5/3}$. Применяя соотношения (23), получим, что коллективное значение энергии в n-ной масштабной оболочке должно быть $E_n \sim q^{(-2/3)n}$. В то же время энергия n-ной оболочки в каскадной модели – это $|U_n + B_n|^2/2$.

Если требовать существования в каскадной модели стационарного решения, соответствующего закону Колмогорова, то оно должно иметь вид: $U_n = U_0 q^{-n/3}$ и $_n = B_0 q^{-n/3}$, где U_0 и B_0 – заданные числа.

Спектральные законы турбулентности выполняются в пределах инерционного интервала масштабов, где нет внешних источников и диссипации, поэтому подставим стационарные решения в уравнения каскадной модели (5) без источника и диссипативных членов. Подстановка даст стационарные уравнения

$$\begin{aligned} &\imath q^n \sum_{i,j} \left(\hat{S}_{ij} |U_0|^2 q^{-(2n+i+j)/3} + \hat{L}_{ij} |B_0|^2 q^{-(2n+i+j)/3} \right) = 0, \\ &\imath q^n \sum_{i,j} \hat{W}_{ij} |U_0 B_0| q^{-(2n+i+j)/3} = 0. \end{aligned} \tag{24}$$

После очевидных преобразований получим уравнения на коэффициенты модели

$$\begin{gathered}\sum_{i,j}\left(\left|\frac{U_0}{B_0}\right|^2 S_{ij}+L_{ij}\right)q^{-(i+j)/3}=0,\\ \sum_{i,j}\hat{W}_{ij}q^{-(i+j)/3}=0.\end{gathered} \tag{25}$$

Подобные уравнения составляются для каждого вида стационарных степенных по масштабам решений.

В отличие от требования существования инвариантов, требование существования степенных стационарных решений не всегда учитывают при расчете коэффициентов модели. Это связано с тем, что существование степенных решений вовсе не гарантирует их устойчивости. Более того, численные эксперименты в некоторых простейших каскадных моделях свидетельствуют о неустойчивости таких решений и реализации степенных распределений только в среднем [1]. Отметим также, что степенное распределение в комплексной каскадной модели связано с модулями коллективных переменных, т.е. надо говорить о стационарных значениях только модулей, аргументы могут при этом меняться. Однако условие существования решений со степенным стационарным распределением модулей переменных приводит к нелинейным уравнениям на коэффициенты, аналитическое решение которых не представляется возможным.

Итак, на выходе блока 2 схемы описываемой технологии получаются линейные однородные уравнения на коэффициенты каскадной модели для каждого спектрального закона, если разработчик этой модели хочет обеспечить их существование.

3.3. Возможность и предельная дальность взаимодействия оболочек. Для того, чтобы волны из оболочек P_n, P_{n+i} и P_{n+j} могли взаимодействовать, надо, чтобы нашлись такие числа $a \in D_n$, $b \in D_{n+i}$ и $c \in D_{n+j}$, что

$$a+b \geqslant c, \quad a+c \geqslant b, \quad b+c \geqslant a. \tag{26}$$

Ясно, что в действительности это не зависит от n, поэтому можно положить $n = 0$ и рассматривать оболочки P_0, P_i и P_j. Поэтому возможность взаимодействия определяется только парой индексов (i, j).

Соответствующие области на плоскости индексов зависят от q, но имеют характерную форму, представленную на рисунке 4. Серым цветом выделены области возможных взаимодействий для $q = \left(1+\sqrt{5}\right)/2$ и $q \geqslant 2$. Области неограниченно продолжаются вдоль осей влево и вниз и по диагонали вправо-вверх с сохранением формы.

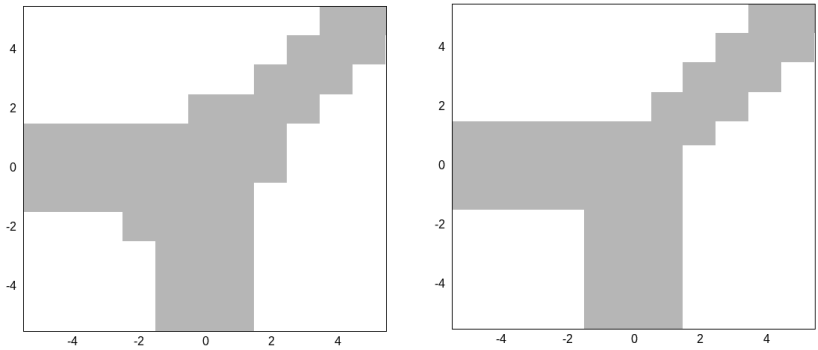


Рис. 4. Области возможных взаимодействий оболочек P_n, P_{n+i} и P_{n+j} (i,j) на плоскости индексов (i,j) для $q = \left(1+\sqrt{5}\right)/2$ и $q = 2$

Пусть предикат $Q(i,j)$ определяет область взаимодействий для заданного q. Например, для $q = 2$ он имеет вид

$$Q = (|i| \leqslant 1 \wedge j \leqslant 1) \vee (|j| \leqslant 1 \wedge i \leqslant 1) \vee (i \geqslant 1 \wedge j \geqslant 1 \wedge |i-j| \leqslant 1)\,.$$

Области возможных взаимодействий на рисунке 4 неограниченны, однако если считать не принципиальную возможность взаимодействия трех выбранных в оболочках волн, а вероятность этого взаимодействия при случайном и равномерном выборе волн в оболочках, то она очень резко спадает с ростом модулей индексов. Такие вероятности легко считаются методом Монте-Карло, подробнее об этом будет сказано ниже.

Тогда маловероятными взаимодействиями можно пренебречь, и в каскадную модель закладывается предельная «дальность» взаимодействия P. Можно сказать, что P – это параметр модели, характеризующий ее нелокальность в пространстве масштабов.

Итак, на выходе блоков 3 и 4 схемы мы получаем выражение для предиката $Q(i,j)$ и число P. Все коэффициенты модели, индексы которых удовлетворяют условию

$$|i| \geqslant P \vee |j| \geqslant P \vee \neg Q(i,j), \tag{27}$$

полагаются нулевыми.

3.4. Составление системы уравнений для коэффициентов модели и ее решение. В блоке 5 схемы на рисунке 1 составляется система уравнений на коэффициенты модели. Эта система включает, прежде всего, семейства уравнений, полученных в блоке 1, причем (19, 20) включаются в любом случае, а из (21) и (22) – только одно в зависимости от размерности моделируемого потока. Формально уравнений в каждом семействе, а значит, и в системе бесконечно много, однако с учетом ограничений на предельную дальность взаимодействия P только конечное число уравнения не выродится в тождества.

Понятно, что составление самой системы и ее последующее решение – задача трудоемкая. Однако ее можно программно реализовать с помощью CAS.

Прежде всего, отметим, что в каждой их групп уравнений (19, 20, 21, 22) пара индексов (i,j) однозначно определяет уравнение в пределах группы – является его идентификатором. Само уравнение может содержать еще $\pm(i-j)$, $-i$ и $-j$ в качестве индексов искомых коэффициентов. Если все эти индексы по модулю окажутся больше P, т.е.

$$(|i| > P \vee |j| > P) \wedge (|i| > P \vee |j-i| > P) \wedge (|j| > P \vee |i-j| > P),$$

то уравнение будет тождеством. Тогда не обязательно выродятся в тождества только уравнения с идентификатором (i,j), задаваемым предикатом

$$\begin{aligned} d(i,j) = {} & (|i| \leqslant P \wedge |j| \leqslant P) \vee (|i| \leqslant P \wedge |i-j| \leqslant P) \vee \\ & \vee (|j| \leqslant P \wedge |i-j| \leqslant P). \end{aligned} \tag{28}$$

Геометрически этот предикат выделяет на плоскости (i,j) звездообразную область D, изображенную на рисунке 5. Отметим, что и среди этих уравнений некоторые могут оказаться тождествами.

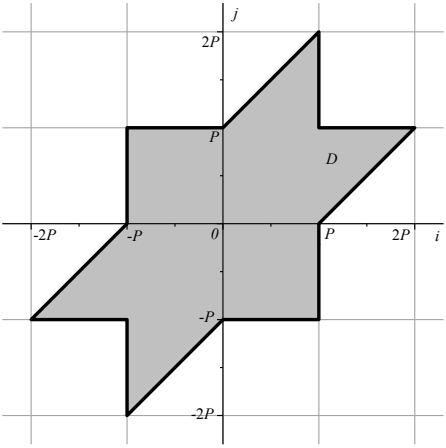


Рис. 5. Область пар индексов (i, j), которые определяют уравнения, не обязательно вырождающиеся в тождества. Пояснения в тексте

Тогда содержание блока 5 схемы заключается в следующих этапах:

1. Объявляем двумерные массивы коэффициентов $\hat{S}_{i,j}$, $\hat{L}_{i,j}$, $\hat{W}_{i,j}$ с индексами $i, j = -2P, \ldots, 2P$.

2. Для индексов $i, j = -2P, \ldots, 2P$ и не удовлетворяющих условию (27) элементам массивов присваиваем нулевые значения. Остальные элементы остаются неопределенными, но с ограничениями $\hat{S}_{i,j} = \hat{S}_{j,i}$ и $\hat{L}_{i,j} = \hat{L}_{j,i}$.

3. Объявляем список левых частей уравнений на коэффициенты и полагаем его пустым.

4. Для $i, j = -2P, \ldots, 2P$, если (i, j) удовлетворяют условию (28), формируем выражение левой части каждого из уравнений, обеспечивающих инварианты (выход блока 1). Если полученное выражение ненулевое, формируем уравнение и добавляем его в список.

5. Если необходимо обеспечить существование стационарных степенных по масштабу решений, формируем уравнения вида (25) (выход блока 2) и добавляем их в список.

6. На основе списка составляем однородную систему линейных алгебраических уравнений.

Понятно, что эти этапы легко программно реализовать средствами любой CAS. Приведем псевдокод составления этой системы уравнений для коэффициентов (Листинг 3).

```
k := 1;
for i := −2P to 2P do for j := −2P to i do
  if Q(i, j) then
    Ŝ_{i,j} := x_k; Ŝ_{j,i} := x_k; k := k + 1;
    L̂_{i,j} := x_k; L̂_{j,i} := x_k; k := k + 1;
  else
    Ŝ_{i,j} := 0; L̂_{i,j} := 0;
  end if;
end do; end do;
for i := −2P to 2P do for j := −2P to 2P do
  if Q(i, j) then
    Ŵ_{i,j} := x_k; k := k + 1;
  else
    Ŵ_{i,j} := 0;
  end if;
end do; end do;
num_var := k − 1;
n := 1;
list_eq := [ ];
for l := −2P to 2P do for m := −2P to 2P do
  if d(l, m) then
    for expr_eq ∈ L_inv do
      expr := Subs(expr_eq, ['i' = l,' j' = m]);
      if expr ≠ 0 then
        list_eq := list_eq ∪ [expr = 0];
        n := n + 1;
      end if;
    end do;
  end if;
end do; end do
for expr_eq ∈ L_stat do
  expr_sum := 0;
  for l = −P to P do for m = −P to P do
    expr := Subs(expr_eq, ['i' = l,' j' = m]);
    if expr ≠ 0 then
      expr_sum := expr_sum + expr;
    end if;
  end do; end do;
  if expr_sum ≠ 0 then
    list_eq := list_eq ∪ [expr_sum = 0];
    n := n + 1;
  end if;
end do;
num_eq := n − 1;
```

Листинг 3. Псевдокод составления системы уравнений для коэффициентов модели

Предполагаем, что выражения левых частей уравнений (19, 20) и (21) или (22) с неопределенными значениями индексов i, j и параметра q предварительно записаны в списке L_inv, а выражения (i,j)-слагаемых левых частей уравнений (25) записаны в списке L_stat. В этом псевдокоде функция $Subs$ выполняет подстановку числовых индексов l и m в выражение $expr\_eq$ вместо символов i и j . Соответствующая функция есть в любой CAS.

В результате выполнения этого псевдокода в списке $list\_eq$ сформирована система из num_eq однородных линейных уравнений с вектором неизвестных $\mathbf{x}$, содержащим num_var компонент. Полученная система является выходом блока 5 технологии.

3.5. Параметрический класс моделей. Полученная на выходе блока 5 система может быть аналитически решена средствами CAS, поскольку любая из них содержит солверы для решения линейных систем, которые реализуют ту или иную форму метода Гаусса. На основе найденных элементов вектора решений можно далее получить значения коэффициентов. Приведем соответствующий псевдокод (Листинг 4), где через $linear\_solv$ обозначен солвер CAS для линейных систем.

```
y := linear_solv(list_eq, x);
k := 1;
for i := -2P to 2P do for j := -2P to i do
  if Q(i,j) then
    Ŝ_{i,j} := y_k; Ŝ_{j,i} := y_k; k := k + 1;
    L̂_{i,j} := y_k; L̂_{j,i} := y_k; k := k + 1;
  else
    Ŝ_{i,j} := 0; L̂_{i,j} := 0;
  end if;
end do; end do;
for i := -2P to 2P do for j := -2P to 2P do
  if Q(i,j) then
    Ŵ_{i,j} := y_k; k := k + 1;
  else
    Ŵ_{i,j} := 0;
  end if;
end do; end do;
```

Листинг 4. Псевдокод нахождения значений коэффициентов в уравнениях модели

Поскольку решаемая система однородная, то возможны два случая. В первом система имеет только нулевое решение. Это говорит о том, что каскадной модели с заданным дальнодействием P, заданным набором инвариантов и стационарных решений не существует. Во втором случае

система имеет линейное многообразие решений, и CAS даст выражение для коэффициентов каскадной модели через свободные параметры. Эти выражения будут линейными однородными по параметрам и дробно-рациональными по q. После выполнения псевдокода Листинга 4 в массивах $\hat{S}_{i,j}$, $\hat{L}_{i,j}$, $\hat{W}_{i,j}$ будут записаны выражения для коэффициентов (выход блока 6 схемы). В результате получим параметрический класс моделей (выход блока 7), возможно, пустой, если модели не существует.

Если разработчик модели не предполагает получение числовых значений коэффициентов, то вычисления на этом заканчиваются (блок 8). В противном случае реализуются этапы блоков 9-11.

3.6. Согласование коэффициентов нелинейных членов и вероятностей взаимодействия оболочек. Ранее в работе уже упоминалось, что вероятности взаимодействия случайно выбранных волн из трех оболочек P_n, P_{n+i}, P_{n+j} сильно отличаются для различных i и j и не зависят от n. Ясно также, что они симметричны по i и j. Эти вероятности легко посчитать методом Монте-Карло для заданного числового значения q в большой серии испытаний. В каждом испытании случайно выбираются три числа, равномерно распределенные в масштабных диапазонах D_0, D_i, D_j. Если эти три числа удовлетворяют неравенствам треугольника, испытание считается успешным. В качестве оценок вероятностей принимаются относительные частоты успешных испытаний.

На рисунке 6 совмещены область возможных взаимодействий и карта вероятностей этих взаимодействий p_{ij} для $q = \left(1+\sqrt{5}\right)/2$. Вероятности вычислялись по 10^6 испытаний для каждой пары (i, j).

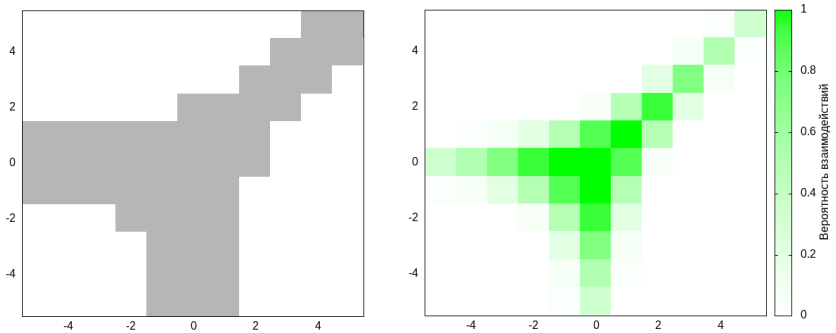


Рис. 6. Область возможных взаимодействий (слева) и карта вероятностей взаимодействия p_{ij} (справа) для $q = \left(1+\sqrt{5}\right)/2$

Псевдокод расчета этих вероятностей приведен в Листинге 5, где через $rnd(\alpha, \beta)$ обозначена стандартная функция генерации равномерно распределенного на отрезке $[\alpha; \beta]$ псевдослучайного числа.

```
q := (1 + √5)/2;
for i := −P to P do
  b_left = q^{−i−1}; b_right = q^{−i};
  for j := −P to i do
    c_left = q^{−j−1}; c_right = q^{−j};
    p_ij := 0;
    for n := 1 to 10^6 do
      a := rnd(q^{−1}, 1);
      b := rnd(b_left, b_right);
      c := rnd(c_left, c_right)
      if a + b ⩾ c and a + c ⩾ b and b + c ⩾ a then p_ij := p_ij + 1;
    end do;
    p_ij := p_ij/10^6;
    p_ji := p_ij;
  end do;
end do;
```

Листинг 5. Псевдокод рачета вероятностей взаимодействия оболочек

Итак, на выходе блока 9 схемы получаем распределение вероятностей взаимодействия масштабных оболочек.

В каскадных моделях (5) коэффициенты при каждом нелинейном члене по абсолютной величине имеют смысл интенсивности взаимодействия волн из оболочек P_n, P_{n+i} и P_{n+j} в процессе, соответствующем этому члену. Например, $|\hat{W}_{ij}|$ характеризует, насколько эффективно генерируются магнитные волны из оболочки P_n механизмом динамо при взаимодействии волн скорости из P_{n+i} с магнитными волнами из P_{n+j}. Поэтому разумно потребовать при расчете коэффициентов модели, чтобы их модули были в каком-то смысле согласованы с вероятностями взаимодействия оболочек.

На выходе блока 7 получены выражения для коэффициентов модели, линейно зависящие от вектора свободных параметров $\mathbf{s} = (s_1, s_2, \ldots, s_m)$. Если какой-либо из коэффициентов оказался равен нулю, это означает, что взаимодействие оболочек именно в этом члене запрещено законами сохранения и согласовывать уже ничего не надо. Поэтому будем рассматривать здесь только ненулевые коэффициенты, зависящие от $\mathbf{s}$.

Ясно, что минимизация различий между модулями коэффициентов и вероятностями p_{ij} требует минимизации выражений вида

$$\left|\frac{\left|\hat{S}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right|, \quad \left|\frac{\left|\hat{L}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right|, \quad \left|\frac{\left|\hat{W}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right|. \tag{29}$$

Нормирование каждой разности в (29) на вероятность p_{ij} уравновешивает большие и малые значения вероятностей в задаче минимизации. Эта минимизация должна быть по всем коэффициентам и является многокритериальной, поэтому допускает различные варианты постановки.

Один из наиболее естественных вариантов постановки с равноправием всех коэффициентов – это минимизация критерия

$$\sum_{i,j}\left[\left|\frac{\left|\hat{S}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right| + \left|\frac{\left|\hat{L}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right| + \left|\frac{\left|\hat{W}_{ij}(\mathbf{s})\right| - p_{ij}}{p_{ij}}\right|\right], \tag{30}$$

которая является содержанием блока 10.

Задача минимизации (30) решалась уже с помощью специально разработанного в C++ приложения, псевдокод которого приведен ниже. Это связано с тем, что вероятности взаимодействий определены с помощью имитационного моделирования, поэтому проведение точных вычислений компьютерной алгебры становится уже бессмысленным. Кроме того, вычисления с плавающей точкой в системах компьютерной алгебры очень медленные, а задача минимизации, представленная выше, требует очень большого объема вычислений.

С другой стороны, хотелось остаться в рамках рациональных значений коэффициентов модели, причем с небольшими числителями и знаменателями. Поэтому использовался перебор рациональных значений свободных параметров вида p/r, где p изменялось в диапазоне $-10, \ldots, 10$, а r в диапазоне $1, \ldots, 10$ с переходом к несократимой форме. Всего таких различных дробей 127, т.е. общее число вариантов свободных параметров 127^m.

Рассмотрим псевдокод этого приложения (Листинг 6). Здесь $next\_placement(\mathbf{z}, n, k)$ означает функцию, которая данное k-элементное размещение с повторениями $\mathbf{z}$ из чисел $1, 2, \ldots, n$ преобразует в непосредственно следующее за ним в лексикографическом порядке и возвращает $true$ [30]. Если такого размещения нет, то она ничего не

изменяет в $\mathbf{z}$ и возвращает $false$. Предполагается, что все 127 возможных значений p/r записаны в векторе $\mathbf{w}$.

```
n := 127;
for l := 1 to m do z_l := 1;
first_placement := true;
do
  for l := 1 to m do s_l := w_{z_l};
  target := 0;
  for i = -P to P do for j = -P to P do
    if Ŝ_ij(s) ≠ 0 then
      target := target + | |Ŝ_ij(s)| - p_ij | / p_ij;
    end if;
    if L̂_ij(s) ≠ 0 then
      target := target + | |L̂_ij(s)| - p_ij | / p_ij;
    end if;
    if Ŵ_ij(s) ≠ 0 then
      target := target + | |Ŵ_ij(s)| - p_ij | / p_ij;
    end if;
  end do; end do;
  if first_placemet then
    min_target := target;
    s_min := s;
    first_placement := false;
  else
    if target < min_target then
      min_target := target;
      s_min := s;
    end if;
  end if;
while next_placement(z, n, m);
```

Листинг 6. Псевдокод для минимизаци критерия (30)

В результате выполнения этого псевдокода в векторе $\mathbf{s}_{min}$ записаны рациональные значения параметров приведенного выше типа, минимизирующие (30).

Рассмотрим некоторые результаты расчетов для $P = 3$ (минимальная нелокальность). На рисунке 7 представлены логарифмы отношения модулей полученных коэффициентов взаимодействия к вероятностям взаимодействия для модели двумерного потока, а на рисунке 8 – для трехмерного потока.

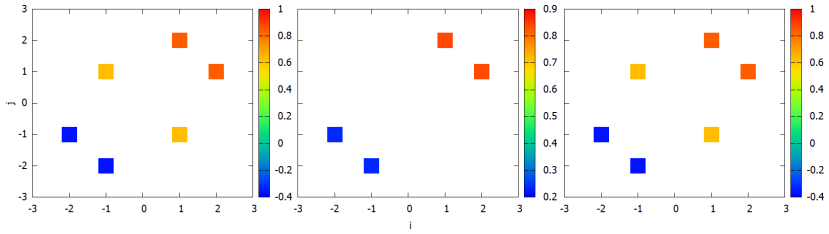


Рис. 7. Отношения моделей полученных коэффициентов взаимодействия к вероятностям взаимодействия в логарифмическом масштабе (двумерный поток). Слева направо коэффициенты $\hat{S}_{ij}$, $\hat{L}_{ij}$, $\hat{W}_{ij}$

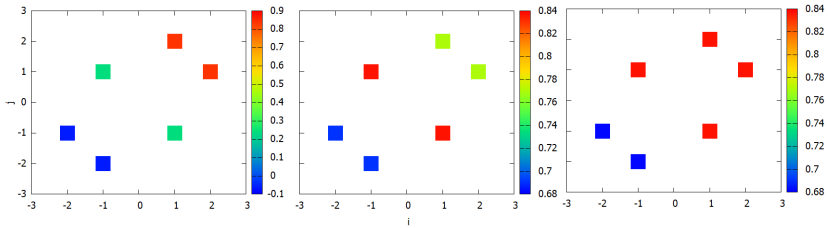


Рис. 8. Отношения полученных коэффициентов взаимодействия к вероятностям взаимодействия в логарифмическом масштабе (трехмерный поток). Слева направо коэффициенты $\hat{S}_{ij}$, $\hat{L}_{ij}$, $\hat{W}_{ij}$

Видно, что модули коэффициентов и вероятности отличаются не более, чем на порядок. Это можно считать хорошим согласованием.

4. Заключение. Результатом настоящей работы является вычислительная технология автоматизированного составления каскадных моделей типа GOY для магнитогидродинамической турбулентности. С ее помощью можно получать уравнения новых каскадных моделей турбулентности, в том числе и нелокальных, с параметрическими выражениями для коэффициентов или с числовыми коэффициентами. В обоих случая сохранение необходимых квадратичных инвариантов и стационарность заданных степенных решений будут абсолютно точными. Разработанная технология может быть полезным вычислительным инструментом для специалистов, занимающимся моделированием мелкомасштабной турбулентности.

Идея технологии и ее реализация принадлежат авторам работы. Технология реализована с использованием систем компьютерной алгебры. Они позволяют программно генерировать системы уравнений на коэффициенты составляемых моделей и решать эти системы в

общем формульном виде. В результате получаются параметрические классы моделей, которые обладают необходимыми квадратичными инвариантами и степенными стационарными решениями. При составлении моделей есть возможность произвольно выбирать предельное дальнодействие в пространстве масштабов – масштабную нелокальность модели. Для последующего численного исследования моделей необходимо фиксировать значения свободных параметров. В предложенной технологии реализован один формальный способ нахождения рациональных значений свободных параметров, при которых коэффициенты нелинейных взаимодействий масштабов будут согласованы с вероятностями этих взаимодействий.

Авторы планируют в дальнейшем использовать технологию для сопряжения крупномасштабных моделей гидромагнитного динамо, описывающих крупные турбулентные вихри, и каскадных моделей для описания мелкомасштабной динамики.

Литература

1. Фрик П.Г. Турбулентность: подходы и модели. Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2010. 332 с.
2. Ditlevsen P.D. Turbulence and shell models. New York: University Press, 2011. 152 p. DOI: 10.1017/CBO9780511919251.
3. Gibbon J.D., Vincenzi D. How to extract a spectrum from hydrodynamic equations? // Journal of Nonlinear Science. 2022. vol. 32. no. 6. pp. 1–25.
4. Gurcan D., Xu S., Morel P. Spiral chain models of two-dimensional turbulence // Physical Review E. 2019. vol. 100. no. 4. DOI: 10.1103/PhysRevE.100.043113.
5. Mailybaev A.A. Hidden scale invariance of intermittent turbulence in a shell model // Physical Review Fluids. 2021. vol. 6. no. 1. DOI: 10.1103/PhysRevFluids.6.L012601.
6. Plunian, F., Stepanov R., Frick P. Shell models of magnetohydroodynamic turbulence // Physics Reports. 2013. vol. 523. no. 1. pp. 1–60.
7. Munoz V., Dominguez M., Riquelme M., Nigro G., Carbone V. Fractality of an mhd shell model for turbulent plasma driven by solar wind data: a review // Journal of Atmospheric and Solar-Terrestrial Physics. 2021. vol. 214. DOI: 10.1016/j.jastp.2020.105524.
8. Chen N., Li Y., Lunasin E. An efficient continuous data assimilation algorithm for the sabra shell model of turbulence // Chaos. 2021. vol. 31. no. 10. DOI: 10.1063/5.0057421.
9. Li L., Liu P., Xing Y., Guo H. Shell models for confined rayleigh–taylor turbulent convection // Communications in Nonlinear Science and Numerical Simulation. 2020. vol. 84. DOI: 10.1016/j.cnsns.2020.105204.
10. Verdini A., Grappin R., Montagud-Camps V. Turbulent heating in the accelerating region using a multishell model // Solar Physics. 2019. vol. 294. DOI: 10.1007/s11207-019-1458-y.
11. Bhadra A., Mishra P.K. Energy spectrum and energy budget of superfluid turbulence using two-fluid shell model // AIP Advances. 2022. vol. 12. no. 2. DOI: 10.1063/5.0083847.
12. Nabil H., Balhamri A., Belafhal A. Propagation of bessel-gaussian shell-model beam through a jet engine exhaust turbulence // Optical and Quantum Electronics. 2022. vol. 54. no. 6. DOI: 10.1007/s11082-022-03743-3.

13. Tropina A.A., Miles R.B. Parametrical study of aero-optical effects using shell models of turbulence // AIAA Science and Technology Forum and Exposition, AIAA SciTech Forum 2022. DOI: 10.2514/6.2022-0986.
14. Inage S. Control parameter optimization for turbulence shell model // Computers and Fluids. 2021. vol. 229. DOI: 10.1016/j.compfluid.2021.105084.
15. Mailybaev A.A. Solvable intermittent shell model of turbulence // Communications in Mathematical Physics. 2021. vol. 388. no. 1. pp. 469–478.
16. Gurcan O.D. Dynamical network models of the turbulent cascade // Physica D: Nonlinear Phenomena. 2021. vol. 426. DOI: 10.1016/j.physd.2021.132983.
17. Водинчар Г.М., Фещенко Л.К. Автоматизированная генерация каскадных моделей турбулентности методами компьютерной алгебры. // Вычислительные технологии. 2021. Т. 26. № 5. С. 65–80.
18. Водинчар Г.М., Фещенко Л.К., Подлесный Н.В. Генерация комплексных каскадных моделей турбулентных систем методами компьютерной алгебры // Вестник КРАУНЦ. Физико-математические науки. 2022. Т. 41. № 4. С. 9–31.
19. Vodinchar G.M., Feshchenko L.K. Computational Technology for the Basis and Coefficients of Geodynamo Spectral Models in the Maple System // Mathematics. 2023. vol. 11(13). DOI: 10.3390/math11133000.
20. Водинчар Г.М., Фещенко Л.К. Применение компьютерной алгебры для составления спектральных моделей кинематического осесимметричного динамо // Вычислительные технологии. 2023. Т. 28. № 2. С. 4–18.
21. Bright C., Kotsireas I., Ganesh V. Applying computer algebra systems with SAT solvers to the Williamson conjecture // Jour. Symbolic Comp. 2020. vol. 100. pp. 187–209.
22. Gayoso Martinez, V., Hernandez Encinas, L., Martin Munoz, A., Queiruga Dios, A. Using Free Mathimatical Software in Engineering Classes // Axioms. 2021. vol. 10(4). DOI: 10.3390/axioms10040253.
23. Bazan E.R., Hubert E. Multivariate interpolation: Preserving and exploiting symmetry // Journal of Symbolic Computation. 2021. vol. 107. pp. 1–22.
24. Conceicao A.C., Pires J.C. Symbolic Computation Applied to Cauchy Type Singular Integrals // Math. Comput. Appl. 2022. vol. 27(1). DOI: 10.3390/mca27010003.
25. Campo-Montalvo E., Fernandez de Sevilla M., Magdalena Benedito J.R., Perez-Diaz S. Some New Symbolic Algorithms for the Computation of Generalized Asymptotes // Symmetry. 2023. vol. 15. no. 1. DOI: 10.3390/sym15010069.
26. Кирсанов М.Н. Математика и программирование в Maple. М.: Ай Пи Ар Медиа, 2020. 160 с.
27. Wang F.Y. Physics with Maple: The Computer Algebra Resource for Mathematical Methods in Physics. New York: Wiley, 2006. 625 p.
28. Campanelli L. One-dimensional model of freely decaying two-dimensional turbulence // Journal of the Korean Physical Society. 2022. vol. 80. no. 10. pp. 972–980.
29. Campanelli L. Dimensional analysis of two-dimensional turbulence // Modern Physics Letters B. 2019. vol. 33. no. 19. DOI: 10.1142/S021798491950218X.
30. Федоряева Т.И. Комбинаторные алгоритмы. Новосибирск: НГУ, 2011. 118 с.

Водинчар Глеб Михайлович — канд. физ.-мат. наук, доцент, ведущий научный сотрудник, и.о. заведующего лабораторией, лаборатория моделирования физических процессов, Институт космофизических исследований и распространения радиоволн ДВО РАН. Область научных интересов: моделирование гидромагнитного динамо, динамические системы с памятью, нелинейная динамика, методы компьютерной алгебры в геофизике и магнитогидродинамике. Число научных публикаций — 140. gvodinchar@ikir.ru; улица

Мирная, 7, 684034, село Паратунка, Елизовский район, Камчатский край, Россия; р.т.: +7(41531)33-193.

Фещенко Любовь Константиновна — канд. физ.-мат. наук, научный сотрудник, лаборатория моделирования физических процессов, Институт космофизических исследований и распространения радиоволн ДВО РАН. Область научных интересов: математическое моделирование в магнитогидродинамических задачах, применение систем символьных вычислений в магнитогидродинамике, каскадные модели, динамические системы, геодинамо. Число научных публикаций — 57. feshenko.lk@yandex.ru; улица Мирная, 7, 684034, село Паратунка, Елизовский район, Камчатский край, Россия; р.т.: +7(41531)33-193.

Поддержка исследований. Работа выполнена за счет Государственного задания ИКИР ДВО РАН (рег. № темы 124012300245-2).

DOI 10.15622/ia.23.6.4

G. VODINCHAR, L. FESHCHENKO

COMPUTATIONAL TECHNOLOGY FOR SHELL MODELS OF MAGNETOHYDRODYNAMIC TURBULENCE CONSTRUCTING

Vodinchar G., Feshchenko L. **Computational Technology for Shell Models of Magnetohydrodynamic Turbulence Constructing.**

Abstract. The paper discusses the computational technology for constructing one type of small-scale magnetohydrodynamic turbulence models – shell models. Any such model is a system of ordinary quadratic nonlinear differential equations with constant coefficients. Each phase variable is interpreted in absolute value as a measure of the intensity of one of the fields of the turbulent system in a certain range of spatial scales (scale shell). The equations of any shell model must have several quadratic invariants, which are analogues of conservation laws in ideal magnetohydrodynamics. The derivation of the model equations consists in obtaining such expressions for constant coefficients for which the predetermined quadratic expressions will indeed be invariants. Derivation of these expressions «manually» is quite cumbersome and the likelihood of errors in formula transformations is high. This is especially true for non-local models in which large-scale shells that are distant in size can interact. The novelty and originality of the work lie in the fact that the authors proposed a computational technology that allows one to automate the process of deriving equations for shell models. The technology was implemented using computer algebra methods, which made it possible to obtain parametric classes of models in which the invariance of given quadratic forms is carried out absolutely accurately – in formula form. The determination of the parameter values in the resulting parametric class of models is further carried out by agreement with the measures of the interaction of shells in the model with the probabilities of their interaction in a real physical system. The idea of the described technology and its implementation belong to the authors. Some of its elements were published by the authors earlier, but in this work, for the first time, its systematic description is given for models with complex phase variables and agreement of measures of interaction of shells with probabilities. There have been no similar works by other authors previously. The technology allows you to quickly and accurately generate equations for new non-local turbulence shell models and can be useful to researchers involved in modeling turbulent systems simulating.

Keywords: automation of modeling, computer algebra, turbulence, magnetohydrodynamics, shell models.

References

1. Frick P.G. Turbulentnost': podhody i modeli [Turbulence: Approaches and Models]. Moscow-Izhevsk: NRC «Regular and Chaotic Dynamic», 2010. 342 p. (In Russ.).
2. Ditlevsen P.D. Turbulence and shell models. New York: University Press, 2011. 152 p. DOI: 10.1017/CBO9780511919251.
3. Gibbon J.D., Vincenzi D. How to extract a spectrum from hydrodynamic equations? Journal of Nonlinear Science. 2022. vol. 32. no. 6. pp. 1–25.
4. Gurcan D., Xu S., Morel P. Spiral chain models of two-dimensional turbulence. Physical Review E. 2019. vol. 100. no. 4. DOI: 10.1103/PhysRevE.100.043113.
5. Mailybaev A.A. Hidden scale invariance of intermittent turbulence in a shell model. Physical Review Fluids. 2021. vol. 6. no. 1. DOI: 10.1103/PhysRevFluids.6.L012601.
6. Plunian, F., Stepanov R., Frick P. Shell models of magnetohydroodynamic turbulence. Physics Reports. 2013. vol. 523. no. 1. pp. 1–60.

7. Munoz V., Dominguez M., Riquelme M., Nigro G., Carbone V. Fractality of an mhd shell model for turbulent plasma driven by solar wind data: a review. Journal of Atmospheric and Solar-Terrestrial Physics. 2021. vol. 214. DOI: 10.1016/j.jastp.2020.105524.
8. Chen N., Li Y., Lunasin E. An efficient continuous data assimilation algorithm for the sabra shell model of turbulence. Chaos. 2021. vol. 31. no. 10. DOI: 10.1063/5.0057421.
9. Li L., Liu P., Xing Y., Guo H. Shell models for confined rayleigh–taylor turbulent convection. Communications in Nonlinear Science and Numerical Simulation. 2020. vol. 84. DOI: 10.1016/j.cnsns.2020.105204.
10. Verdini A., Grappin R., Montagud-Camps V. Turbulent heating in the accelerating region using a multishell model. Solar Physics. 2019. vol. 294. DOI: 10.1007/s11207-019-1458-y.
11. Bhadra A., Mishra P.K. Energy spectrum and energy budget of superfluid turbulence using two-fluid shell model. AIP Advances. 2022. vol. 12. no. 2. DOI: 10.1063/5.0083847.
12. Nabil H., Balhamri A., Belafhal A. Propagation of bessel-gaussian shell-model beam through a jet engine exhaust turbulence. Optical and Quantum Electronics. 2022. vol. 54. no. 6. DOI: 10.1007/s11082-022-03743-3.
13. Tropina A.A., Miles R.B. Parametrical study of aero-optical effects using shell models of turbulence. AIAA Science and Technology Forum and Exposition, AIAA SciTech Forum 2022. DOI: 10.2514/6.2022-0986.
14. Inage S. Control parameter optimization for turbulence shell model. Computers and Fluids. 2021. vol. 229. DOI: 10.1016/j.compfluid.2021.105084.
15. Mailybaev A.A. Solvable intermittent shell model of turbulence. Communications in Mathematical Physics. 2021. vol. 388. no. 1. pp. 469–478.
16. Gurcan O.D. Dynamical network models of the turbulent cascade. Physica D: Nonlinear Phenomena. 2021. vol. 426. DOI: 10.1016/j.physd.2021.132983.
17. Vodinchar G.M., Feshenko L.K. [Automated generation of turbulence shell models by the methods of computer algebra]. Vychislitel'nye tehnologii – Computational Technologies. 2021. vol. 26. no. 5. pp. 65–80. (In Russ.).
18. Vodinchar G.M., Feshenko L.K. Podlesnyi N.V. [Construction of complex shell models of turbulent systems by computer algebra methods]. Vestnik KRAUNC. Fiziko-matematicheskie nauki – Vestnik KRAUNC. Physical and mathematical sciences. 2022. vol. 41. no. 4. pp. 9–31. (In Russ.).
19. Vodinchar G.M., Feshchenko L.K. Computational Technology for the Basis and Coefficients of Geodynamo Spectral Models in the Maple System. Mathematics. 2023. vol. 11(13). DOI: 10.3390/math11133000.
20. Vodinchar G.M., Feshenko L.K. [Computer algebra application for the developed of the spectral models of kinematic axisymmetric dynamo]. Vychislitel'nye tehnologii – Computational Technologies. 2023. vol. 28. no. 2. pp. 4–18. (In Russ.).
21. Bright C., Kotsireas I., Ganesh V. Applying computer algebra systems with SAT solvers to the Williamson conjecture. Jour. Symbolic Comp. 2020. vol. 100. pp. 187–209.
22. Gayoso Martinez, V., Hernandez Encinas, L., Martin Munoz, A., Queiruga Dios, A. Using Free Mathimatical Software in Engineering Classes. Axioms. 2021. vol. 10(4). DOI: 10.3390/axioms10040253.
23. Bazan E.R., Hubert E. Multivariate interpolation: Preserving and exploiting symmetry. Journal of Symbolic Computation. 2021. vol. 107. pp. 1–22.
24. Conceicao A.C., Pires J.C. Symbolic Computation Applied to Cauchy Type Singular Integrals. Math. Comput. Appl. 2022. vol. 27(1). DOI: 10.3390/mca27010003.
25. Campo-Montalvo E., Fernandez de Sevilla M., Magdalena Benedito J.R., Perez-Diaz S. Some New Symbolic Algorithms for the Computation of Generalized Asymptotes. Symmetry. 2023. vol. 15. no. 1. DOI: 10.3390/sym15010069.

26. Kirsanov M.N. Matematika i programmirovanie v Maple [Mathematics and programming in Maple]. M.: APR Media, 2020. 160 p. (In Russ.).
27. Wang F.Y. Physics with Maple: The Computer Algebra Resource for Mathematical Methods in Physics. New York: Wiley, 2006. 625 p.
28. Campanelli L. One-dimensional model of freely decaying two-dimensional turbulence. Journal of the Korean Physical Society. 2022. vol. 80. no. 10. pp. 972–980.
29. Campanelli L. Dimensional analysis of two-dimensional turbulence. Modern Physics Letters B. 2019. vol. 33. no. 19. DOI: 10.1142/S021798491950218X.
30. Fedorjaeva T.I. Kombinatornye algoritmy [Combinatorial algorithms] Novosibirsk: NSU, 2011. 118 p. (In Russ.).

Vodinchar Gleb — Ph.D., Associate Professor, Leading researcher, acting head of the laboratory, Laboratory for simulation of physical processes, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS. Research interests: hydromagnetic dynamo simulation, dynamical systems with memory, nonlinear dynamics, computer algebra methods in geophysics and magnetohydrodynamics. The number of publications — 140. gvodinchar@ikir.ru; 7, Mirnaya St., 684034, Paratunka, Elizovskiy district, Kamchatka region, Russia; office phone: +7(41531)33-193.

Feshchenko Lyubov — Ph.D., Researcher, Laboratory for modeling physical processes, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS. Research interests: mathematical modeling in magnetohydrodynamic problems, application of symbolic computation systems in magnetohydrodynamics, cascade models, dynamic systems, geodynamo. The number of publications — 57. feshenko.lk@yandex.ru; 7, Mirnaya St., 684034, Paratunka, Elizovskiy district, Kamchatka region, Russia; office phone: +7(41531)33-193.

Acknowledgements. The work was supported by IKIR FEB RAS State Task (subject registration № 124012300245-2).

УДК 621.391 **DOI** 10.15622/ia.23.6.5

С.В. Дворников, Д.В. Васильева

ПОВЫШЕНИЕ ДОСТОВЕРНОСТИ ВЫЯВЛЕНИЯ АНОМАЛИЙ НА ИЗОБРАЖЕНИЯХ ПРИ ФОРМИРОВАНИИ ИХ ВЕКТОРОВ ПРИЗНАКОВ В БАЗИСАХ ВЕЙВЛЕТОВ

Дворников С.В., Васильева Д.В. **Повышение достоверности выявления аномалий на изображениях при формировании их векторов признаков в базисах вейвлетов.**

Аннотация. Предложен способ обнаружения спасательных плотов и шлюпок в акватории морей и океанов после кораблекрушений, основанный на распознавании аномалий на обрабатываемых изображениях, что увеличивает вероятность распознавания объектов мониторинга. Обоснован подход к решению такой задачи. Представлена постановка задачи распознавания объектов с позиций бинарной классификации при обнаружении аномалий. Получено аналитическое выражение для алгоритма принятия решения. Рассмотрена возможность формализации матриц изображений в виде гистограмм распределений интенсивности цветности (яркости). Оценена контрастность признакового пространства на их основе. Предложено повысить контрастность признаковых пространств за счет вторичной обработки гистограмм распределений в базисе кратномасштабной вейвлет-декомпозиции. Рассмотрена возможность реализации вейвлет-преобразований на основе функций Хаара и вейвлетов Гаусса 1-го и 2-го порядков. Обоснован механизм формирования вторичных векторов признаков из трехмерных вейвлет-преобразований, путем усреднения их коэффициентов по оси временного сдвига. Показано, что при одинаковой размерности гистограмм распределения яркости с вновь формируемыми векторами признаков, последние обеспечивают более высокую контрастность признаковых пространств. Рекомендовано для формализации изображений в формате jpeg использовать вейвлет Гаусса 2-го порядка, обеспечивающий при прочих равных условиях большую величину различий для изображений, содержащих аномалии. Разработан подход к вероятностной оценке алгоритма автоматического распознавания изображений. Получено аналитическое выражение и обоснованы его составляющие элементы. Приведены графические зависимости вероятности правильного обнаружения (распознавания) аномалий, в зависимости от размеров по отношению к общей площади кадра и дисперсии подстилающего фона. Представлены результаты эксперимента по распознаванию изображений со спасательной шлюпкой в акватории океана. Определены направления дальнейших исследований.

Ключевые слова: распознавание изображений, векторы признаков, вейвлет-преобразование, контрастность признаковых пространств.

1. Введение. По данным ООН (https://en.m.wikipedia.org/wiki/List_of_shipwrecks_in_2023. Свободный доступ на 29.04.2024) только за последний год произошло 205 крупных кораблекрушений. Поэтому вопросы проведения аварийно-спасательных работ на воде занимают одно из приоритетных направлений в деятельности Министерства по чрезвычайным ситуациям [1]. В интересах поиска и спасения потерпевших кораблекрушение создаются специальные подразделения,

занимающиеся мониторингом водного пространства [2, 3], в том числе и с использованием беспилотной авиации [4, 5].

Поэтому одной из приоритетных задач является разработка эффективного научно-методического аппарата, позволяющего осуществлять первичную обработку видеоматериалов с целью обнаружения терпящих бедствие на воде.

В настоящее время для этих целей активно используют метод искусственного интеллекта [6, 7]. Однако успешность его применения во многом зависит от качества разработанных алгоритмов идентификации, как правило, базирующихся на элементах теории распознавания образов [8 – 10].

С учетом указанных обстоятельств, в настоящей статье представлен подход к распознаванию объектов с позиций выявления аномалий на обрабатываемых изображениях по отношению к эталонным описаниям. В частности, рассмотрены предложения по повышению контрастности векторов признаков идентифицируемых объектов, за счет их формирования в базисах вейвлет-преобразований [11 – 13].

2. Обоснование подхода к решению задачи распознавания. В общем случае задачу поиска результатов последствий аварий водного транспорта и терпящих бедствие на воде, можно рассматривать с позиций обнаружения аномалий на изображениях, выявляемых в процессе последовательного анализа представляемых для обработки данных [14, 15]. Видеоданные с беспилотного носителя поступают на пост мониторинга, где оператор по результатам их визуального анализа принимает решение о наличии или отсутствии на изображении терпящих бедствие [16]. Но учитывая огромные водные территории, в пределах которых приходится вести поиск, такой подход становится достаточно трудоемким.

В настоящее время широкое применение получили два научных направления, связанных с решением задач обнаружения объектов по результатам обработки изображений.

Первый подход базируется на технологии компьютерного зрения, а в основе другого используются результаты статистической обработки параметров изображений.

Технологии машинного обучения – активно развивающееся научное направление, основанное на применении нейронных сетей. Его сущность базируется на последовательном уточнении исходных алгоритмов, по результатам обработки данных, как на этапе «обучения», так и в процессе работы [17].

Анализ результатов исследований, связанных со сложностью идентификаций различного рода аномалий на водной поверхности, представленных в [18 – 22], показал перспективность применения технологий на основе глубокого обучения нейронных сетей (англ. Deep Neural Networks – DNNs). Среди которых, наилучшие результаты по обнаружению аномалий на водных поверхностях обеспечивают сверточные нейронные сети (англ. CNNs).

По результатам изучения трудностей практической реализации технологий CNNs в рассматриваемой проблематике [21, 22] было установлено, что для обработки изображения одной из наиболее эффективных нейронных сетей SENet необходимо осуществить $5{,}8 \times 10^9$ вычислительных операций. При этом обеспечивается вероятность распознавания на уровне 0,965, но размер обрабатываемого изображения составляет всего лишь 224×224 пикселов.

Это связано с проблемой обучения сети, поскольку результаты и последствия бедствий на воде сложно конкретизировать и затем формализовать в виде поправочных коэффициентов в решающих алгоритмах. Данный вывод сделан в [18].

В свою очередь статистический подход к обработке параметров изображений, базируется на методе классической теории распознавания образов, разработанных К. Фукунагой [23]. Однако его непосредственное применение, в виду высокой вариативности водной поверхности, будет связано со значительными сложностями, что подтверждается результатами, представленными в [24, 25].

Но, несмотря на свою ограниченность [26], статистический подход позволяет реализовать алгоритмы классификации, основанные на выявлении аномалий более простым образом, чем это обеспечат методы машинного обучения.

При этом понятие аномалий будем рассматривать с позиций A. Zimek и E. Schubert, которые в [27] определили данное явление как: «…редкие данные, события или наблюдения, которые вызывают подозрения ввиду их существенного отличия от большей части данных». То есть применительно к рассматриваемой тематике – это объекты на изображениях водных поверхностях, не характерных для акваторий морей и океанов в отсутствии бедствий на воде.

Но для обеспечения эффективности такого подхода необходима разработка соответствующего методического аппарата, позволяющего осуществить требуемую формализацию обрабатываемых изображений до уровня их машинного восприятия и компьютерной обработки [28].

Очевидно, что результатом такой формализации должна быть некая цифровая матрица или вектор числовых значений, которые соответствующим образом отображают содержание обрабатываемого изображения [29, 30]. Только в этом случае возможна реализация параметрического подхода к решению задач классификации с позиций теории распознавания образов.

Таким образом, формализация изображений до уровня данных, позволяющих проводить их цифровую обработку, является основополагающим моментом проводимого исследования.

С учетом рассмотренных моментов выбора подхода к решению задачи обнаружения аномалий в обрабатываемых изображениях, можно заключить, что применение методов искусственного интеллекта в данной ситуации связано с определенными сложностями. Поскольку исполнение такого подхода предполагает необходимость организации высокоскоростного трафика с беспилотной авиацией, находящейся на значительном удалении от центров мониторинга, и наличие дорогостоящего высокопроизводительного оборудования, для реализации алгоритмов идентификации, при том, что существующий уровень техники не в полной мере позволяет эффективно решать задачи такого уровня, в частности, по размерам обрабатываемых изображений.

Поэтому целесообразным видится переход к методам теории распознавания образов, основанных на параметрической обработке фрагментов фото и видео материалов. Однако такой переход предполагает разработку методического аппарата по формализации данных обрабатываемых изображений до уровня, позволяющего формировать на их основе векторы признаков [10]. А учитывая, что эффективность систем распознания определяется контрастностью признакового пространства, основополагающим моментом является поиск подходов, позволяющих указанную контрастность повысить.

В настоящей статье под понятием контрастности будем понимать различия физических признаков, характеризующих объекты (в данном случае изображения), относящихся к альтернативным (распознаваемым) классам [4].

3. Постановка задачи распознавания объектов на изображениях. В терминах методов теории распознавания образов, задачу обнаружения новых объектов на обрабатываемых изображениях можно отнести к задачам классификации [31], согласно которой по результатам обработки каждое изображение относят к одному из альтернативных классов, определяющих признаковое пространство распознавания.

При этом каждый из альтернативных классов будет представлен совокупностью признаков, характеризующих объект, из которых формируются векторы признаков.

Наиболее просто такая задача решается для случая выявления аномалий на обрабатываемых изображениях. То есть когда определяют два альтернативных класса изображений: A_1 – класс изображений на которых отсутствуют аномалии; и A_2 – класс изображений, которые содержат аномалии.

В рамках проводимых исследований, под аномалиями будем понимать наличие на видео и фотоматериалах изображений, относящихся к последствиям происшествий и аварий на воде (обломки судов, спасательные средства, и т.д.).

С позиций методов теории распознавания образов каждый из альтернативных классов характеризуется своими признаками, которые могут быть представлены в виде векторов признаков:

${}^2V_m$ – вектор признаков характеризующий класс A_2; и ${}^1V_m$ – вектор признаков характеризующий класс A_1. Здесь $m = 1, \ldots M$ – текущий параметр, определяющий номер и позицию конкретного признака в векторе признаков, размерностью M [28].

Учитывая возможную вариативность признаков, характеризующих классы A_1 и A_2, для них формируют, так называемые эталонные описания, на основе которых создают эталонные векторы признаков этих классов.

Под эталонными описаниями будем понимать усредненные значения, характеризующие распознаваемый объект (изображение). Причем усреднения могут осуществляться как на уровне построения альтернативных пространств, так и на уровне формирования характеризующих их векторов.

Вместе с тем наличие векторов признаков позволяет свести задачу классификации обрабатываемых текущих изображений к процедурам сравнения их векторов признаков с векторами признаков эталонных описаний альтернативных классов. То есть вектор признаков текущего изображения ${}^0V_m$ будет сравниваться с эталонными векторами признаков двух альтернативных классов A_1 и A_2. Аналитически такая операция представляет собой алгебраическую разность между парами векторов ${}^2V_m$ и ${}^0V_m$, и ${}^1V_m$ и ${}^0V_m$, рассчитываемую как

$$d_{01(02)} = \frac{1}{M}\sum_{m=1}^{M} | \, {}^0V_m - {}^{1(2)}V_m \, |, \qquad (1)$$

где d_{01} – абсолютная разность между вектором обрабатываемого изображения ${}^0V_m$ и вектором ${}^1V_m$, эталонного описания класса A_1. d_{02} – разность между вектором ${}^0V_m$ и вектором ${}^2V_m$.

Заметим, что абсолютная разность d_{01} как раз и определяет контрастность признаков. Для бинарной системы классификации данное значение будет характеризовать и контрастность признакового пространства распознавания. Для многоальтернативных алгоритмов контрастность признакового пространства распознавания будет определяться наибольшими различиями между любыми двумя произвольными векторами признаков.

Тогда искомый алгоритм принятия решения о наличии аномалий может быть представлен в виде условия оценки знаковой функции:

$$\operatorname{sign}(d_{01} - d_{02}) = \begin{cases} 0 > A_0 \in A_2; \\ 0 \le A_0 \in A_1. \end{cases} \tag{2}$$

В этом случае алгоритм классификации, с учетом вычисления аналитического значения разности векторов между обрабатываемым изображением и векторами эталонных описаний альтернативных классов, может быть приведен к следующему виду [4, 10]:

$$d_{01(02)} = \frac{1}{M}\sum_{m=1}^{M} \left| {}^0V_m - {}^{1(2)}V_m \right| \left| d_{01} \begin{matrix} A_0 \in A_2 \\ > \\ \le \\ A_0 \in A_1 \end{matrix} d_{02} \right. . \tag{3}$$

В соответствии с алгоритмом (3) эффективность правильной классификации тем выше, чем выше контрастность между векторами ${}^1V_m$ и ${}^2V_m$.

Согласно выражению (3) чем больше различия между векторами признаков, т.е. чем выше контрастность d_{01}, тем проще сделать правильный выбор в пользу того, или иного класса. Другими словами, высокая контрастность изначально обеспечит высокую достоверность верного решения в условиях мешающих факторов, приводящих к снижению различий между векторами признаков, вызванных, например, вариативностью водной поверхности в условиях отсутствия аномалий.

Следовательно, необходим поиск такого функционального базиса формирования векторов признаков, который при минимальном

различии между альтернативными классами, обеспечит максимальные различия между их векторами признаков, т.е. обеспечит наибольшую разность для алгоритма (3).

Поскольку эффективность распознавания напрямую зависит от физических различий между векторами признаков при прочих равных условиях, то, частную вербальную задачу исследования определим в следующей формулировке.

Из возможных подходов к формализации изображений необходимо выбрать тот, который обеспечит наибольшую контрастность, сформированных на его основе векторов признаков. В общем случае, между двумя любыми альтернативными классами.

Далее полагая, что подход к формализации изображения предполагает его декомпозицию в функциональном базисе до уровня числовых значений, то математически частную задачу исследования запишем в следующем виде:

$$F\{V_m\}_M \max_{\|A_1 - A_2\| \to \min} |\, {}^1V_m - {}^2V_m \,|\,. \tag{4}$$

Здесь оператор $F\{*\}$ – обозначение функционального базиса декомпозиции изображения.

Учитывая результаты, полученные в [12 – 15, 28], предлагается в качестве функционального базиса декомпозиции изображений использовать их кратномасштабные преобразования на основе вейвлетов.

Далее будем полагать, что в рамках практических приложений, формализация изображений должна позволить получать векторы признаков удобные для их последующей алгоритмической обработки.

При этом следует понимать, что в ходе классификации используются процедуры не фактического распознавания изображения, а выявления наличия на нем признаков последствий аварий и происшествий на воде.

4. Выбор базиса формализации изображений. В [32] представлены исследования, согласно которым переход от формата изображений RGB (англ. Red, Green, Blue), к формату GS (англ. Gray Scale) ведет к потере информации лишь для цветов имеющих одинаковую интенсивность освещенности.

Для дальнейшей обработки матриц предлагается использовать декоррелирующие преобразования на основе вейвлет-преобразований [33, 34]. Так как матрица полутонового изображения размером 800 × 600 будет содержать 480 тыс. значений.

Учитывая, что матрица изображения представляет собой набор дискретных значений, предлагается использовать дискретные формы вейвлет-преобразования (ДВП) [35, 36].

В общем случае, реализация ДВП представляет собой результат свертки входной реализации обрабатываемого процесса $x[n]$ и формирующего материнского вейвлета $\psi[n]$.

Физически, реализацию ДВП можно рассматривать как результат декомпозиции входного процесса путем последовательной ВЧ-фильтрации $h[2n - k]$, с целью получения детализирующих значений, определяемых как вейвлет-коэффициенты $y_{\text{ВЧ}}[n]$. И НЧ-фильтрации $g[2n - k]$, для формирования аппроксимирующих коэффициентов $y_{\text{НЧ}}[n]$.

$$y_{\text{НЧ}}[n]\sum_{k=0}^{K} x[k]g[2n-k]. \tag{5}$$

$$y_{\text{ВЧ}}[n]\sum_{k=0}^{K} x[k]h[2n-k]. \tag{6}$$

Фильтры, описываемые выражениями (5) и (6) получили название квадратурных зеркальных фильтров.

Последовательная реализация процедур, определяемая выражениями (5) и (6), формирует двумерную матрицу вейвлет-коэффициентов, определяемую как дискретное вейвлет-преобразование [37]. Таким образом, ДВП можно рассматривать как распределение энергии входной реализации обрабатываемого процесса в пространстве масштабных преобразований и временных сдвигов формирующих (порождающих) материнских вейвлетов.

Простейшим вейвлетом, используемым для формирования ДВП, является вейвлет- Хаара, описываемый выражением [36]:

$$\psi_{\text{X}}[n] = \begin{cases} 1, & 0 \le n < 1/2; \\ -1, & 1/2 \le n < 1; \\ 0, & n \notin [0,1). \end{cases} \tag{7}$$

В соответствии с аналитической формой представления (7), элементы входной реализации группируются по 2. И для каждой группы вычисляется их сумма и разность. Указанная группировка осуществляется рекурсивно, с образованием нового уровня

масштабирования в пространстве временных сдвигов [37]. В результате получают 2^{k-1} уровней декомпозиций и одну общую сумму в виде ошибки (остатка) аппроксимации.

На рисунке 1 показано временное представление отцовского $\varphi_X(n)$ и материнского $\psi_X(n)$ вейвлетов.

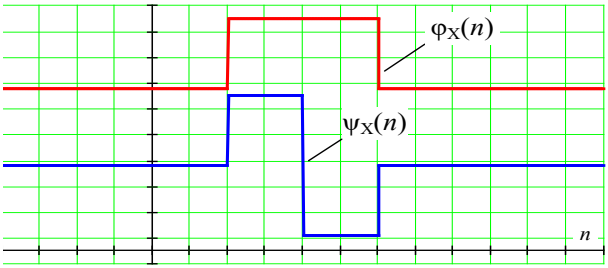


Рис. 1. Временное представление вейвлетов Хаара

Альтернативным решением для ДВП является вейвлет-преобразование на основе дискретных рядов непрерывного времени [38], формируемых, как правило, на основе вейвлетов Гаусса 1-го порядка $\psi_{\Gamma 1}[n]$:

$$\psi_{\Gamma 1}[n] = -n \exp(-n^2/2), \tag{8}$$

и 2-го порядка $\psi_{\Gamma 2}[n]$:

$$\psi_{\Gamma 2}[n] = (1-n^2) \exp(-n^2/2). \tag{9}$$

На рисунке 2 демонстрируются временные структуры вейвлетов Гаусса $\psi_{\Gamma 1}[n]$ и $\psi_{\Gamma 2}[n]$.

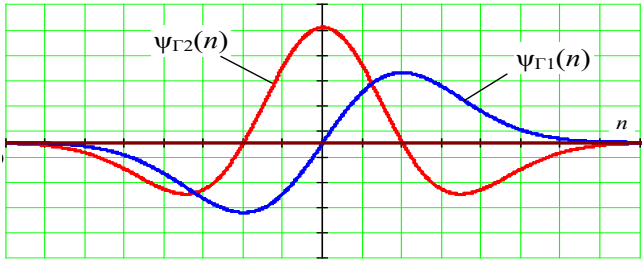


Рис. 2. Временное представление вейвлетов Гаусса

В настоящее время известны в теории и широко применяются на практике большое количество различных типов вейвлет-функций.

Выбор вейвлетов Гаусса на данном этапе обусловлен их хорошими аппроксимирующими свойствами, в приложении к обработке первичных векторов признаков изображений, формируемых в виде гистограмм распределения яркости из матриц полутоновых изображений [28]. В качестве примера представлен кадр изображения водной поверхности в формате *.jpg размером 600×800 пикселей, кодированных 8 битами.

В дальнейшем исследовании данное изображение определено как эталонное. Под понятием эталонного, будем понимать изображение, характеризующее наиболее типовой фон водной поверхности для условий проведения эксперимента (на практике – водная акватория до момента бедствий на воде).

В рамках настоящей статьи, изображение на рисунке 3 определяет класс A_1, т.е. класс, определяющий изображения, а которых отсутствуют аномалии. Для примера, на рисунке 4 изображена матрица данных эталонного изображения.

Рис. 3. Изображение водной поверхности (эталонное)

Структуру матрицы, представленной на рисунке 4, определяют целочисленное значение пиксела в диапазоне от 0 до 255, с координатами по оси абсцисс и ординат, соответствующих ее размерности 600×800. Как уже отмечалось, размеры матрицы существенно затрудняют ее использование для непосредственного применения в алгоритмах распознавания. Поэтому в [10] предложено в качестве вектора признаков использовать гистограмму

распределения значений яркости пикселей, сформированную из матрицы изображений.

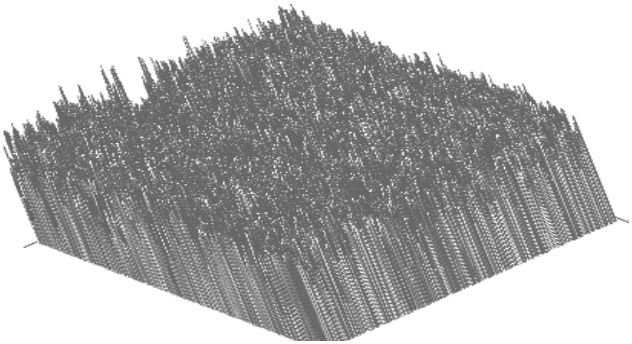


Рис. 4. Матрица данных эталонного изображения водной поверхности

На рисунке 5 демонстрируется гистограмма распределения яркости пикселей эталонного изображения.

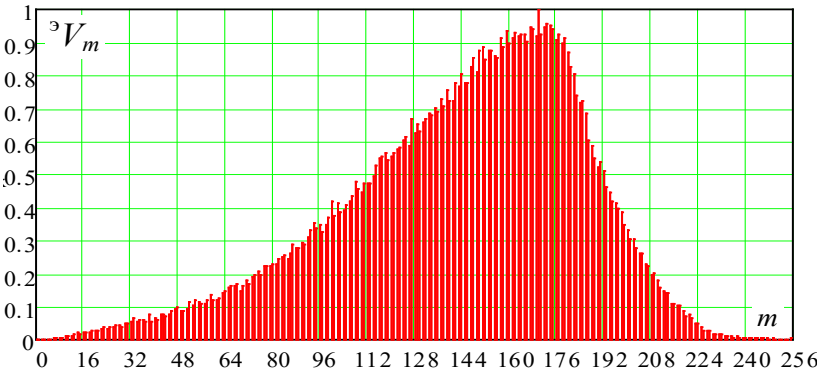


Рис. 5. Гистограмма распределения пикселей яркости эталонного изображения

Поскольку гистограмма распределения яркости представляет собой упорядоченный вектор, количество элементов которого всегда остается неизменным вне зависимости от размера и качества исходного изображения, то открывается возможность использования гистограммы в качестве векторов признаков распознавания.

Для оценки обоснованности такого подхода рассмотрим два кадра морской акватории (рисунок 6 и рисунок 7), на одном из которых изображена спасательная шлюпка.

Рис. 6. Шлюпка на водной поверхности (изображение со шлюпкой)

Отметим, что для эксперимента было отобрано 200 кадров с изображением морской поверхности в пределах одной акватории, полученных примерно в одно и то же время, и одних и тех же погодных условиях.

Рис. 7. Изображение водной поверхности (текущее)

На рисунке 8 и 9, соответственно показаны гистограммы распределения нормированных значений яркости, рассматриваемые в качестве векторов признаков, указанных изображений.

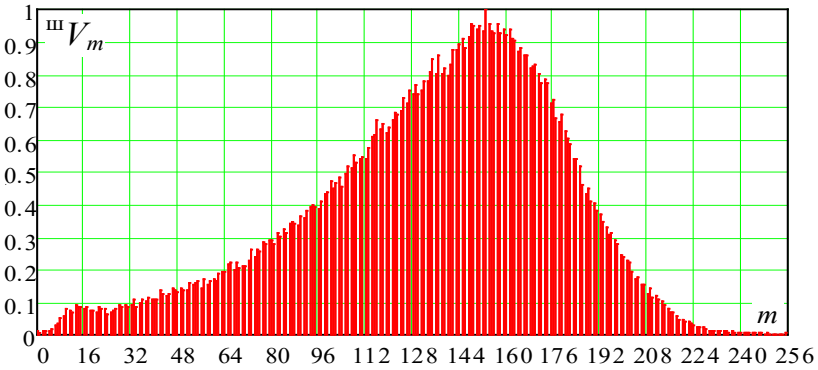


Рис. 8. Гистограмма распределения пикселей яркости изображения со шлюпкой

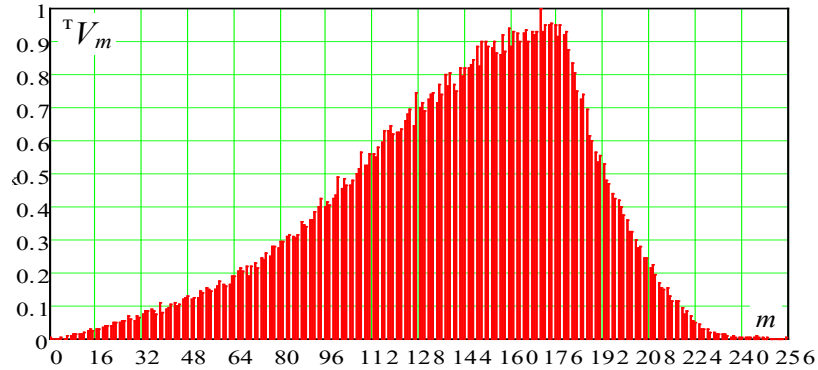


Рис. 9. Гистограмма распределения пикселей яркости текущего изображения

Далее, в соответствии с выражением (1), рассчитанная абсолютная разность между вектором обрабатываемого изображения ${}^{\text{т}}V_m$ и эталонного описания ${}^{\text{э}}V_m$ будет равна

$$ {}^{\text{эт}}d = \frac{1}{256}\sum_{m=0}^{255} | {}^{\text{э}}V_m - {}^{\text{т}}V_m | = 0,03 . $$

А между векторами ${}^{\text{ш}}V_m$ и ${}^{\text{э}}V_m$, соответственно

$$^{эт}d = \frac{1}{256}\sum_{m=0}^{255} | {}^{э}V_m - {}^{т}V_m | = 0,068 .$$

Полученный результат позволяет сделать заключение о потенциальной возможности использования распределения гистограмм пикселей яркости в качестве векторов признаков для распознавания изображений.

Однако более глубокий анализ исследуемых изображений показал, что положительный результат был обеспечен лишь благодаря более низкой дисперсии различий между векторами ${}^{т}V_m$ и ${}^{э}V_m$, ($^{эт}R = 5{,}36\times10^{-3}$) по отношению к дисперсии разности между векторами ${}^{ш}V_m$ и ${}^{э}V_m$ ($^{шэ}R = 7{,}04\times10^{-3}$).

В данном исследовании дисперсия различий разности между векторами рассчитывалась по следующей формуле

$$^{01}R = \frac{1}{M-1}\times \times\sum_{m=0}^{M-1}\left[({}^{0}V_m - {}^{1}V_m) - \frac{1}{M}\sum_{m=0}^{M-1} ({}^{0}V_m - {}^{1}V_m) \right]^2 . \quad (10)$$

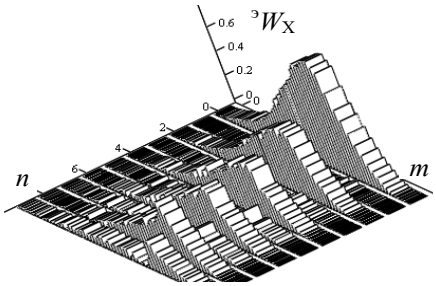
Низкая дисперсия различий между векторами ${}^{т}V_m$ и ${}^{э}V_m$ обусловлена тем, что волнение моря не превышает 1-2 баллов. Очевидно, что при более сильном волнении величина дисперсии, характеризующей различия векторов, возрастет, что существенно снизит вероятность правильной классификации.

Следовательно, необходим поиск подходов, позволяющих повысить контрастность признакового пространства распознавания, определяемого величиной различий векторов признаков, характеризующего его классов.

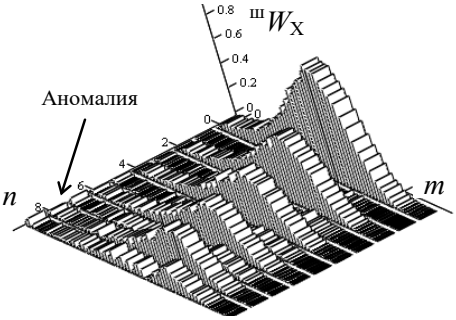
5. Результаты эксперимента. В соответствии с целевой установкой, для повышения контрастности признакового пространства предложено векторы признаков изображений строить на основе ДВП.

Матрицы дискретных значений вейвлет-преобразований для эталонного, текущего и изображения со шлюпкой на основе вейвлета Хаара представлены на рисунке 10.

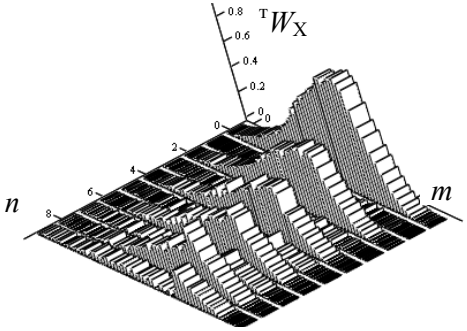
Здесь ${}^{э}W_{\text{X}}$, ${}^{т}W_{\text{X}}$, ${}^{ш}W_{\text{X}}$ – ДВП эталонного, текущего и изображения со шлюпкой на основе вейвлета Хаара.



а)



б)



в)

Рис. 10. а) матрицы эталонного изображения на основе вейвлета Хаара; б) матрицы изображения со шлюпкой на основе вейвлета Хаара; в) матрицы текущего изображения на основе вейвлета Хаара

Матрицы дискретных значений вейвлет-преобразований для эталонного, текущего и изображения со шлюпкой на основе вейвлета Гаусса 1-го порядка представлены на рисунке 11.

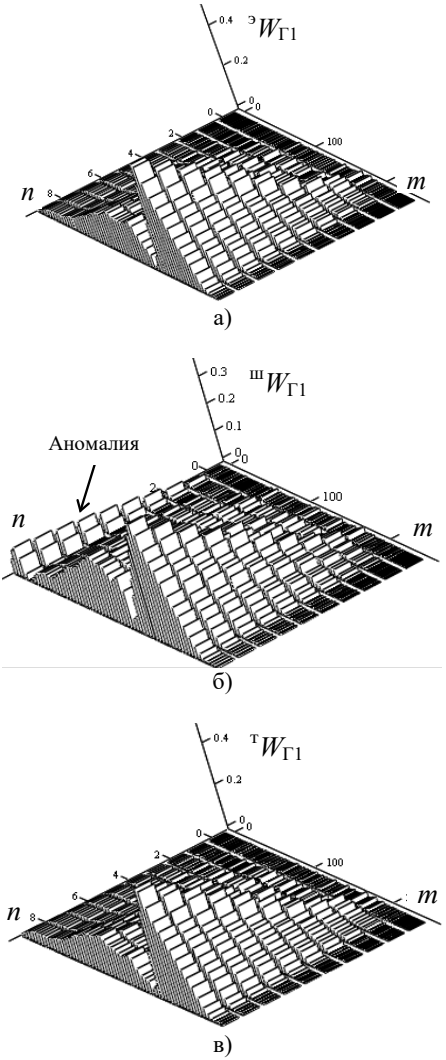


Рис. 11. а) матрицы эталонного изображения на основе вейвлета Гаусса 1-го порядка; б) матрицы изображения со шлюпкой на основе вейвлета Гаусса 1-го порядка; в) матрицы текущего изображения на основе вейвлета Гаусса 1-го порядка

Здесь ${}^{э}W_{Г1}$, ${}^{т}W_{Г1}$, ${}^{ш}W_{Г1}$ – ДВП эталонного изображения на основе Гаусса 1-го порядка.

Матрицы дискретных значений вейвлет-преобразований для эталонного, текущего и изображения со шлюпкой на основе вейвлета Гаусса 2-го порядка представлены на рисунке 12.

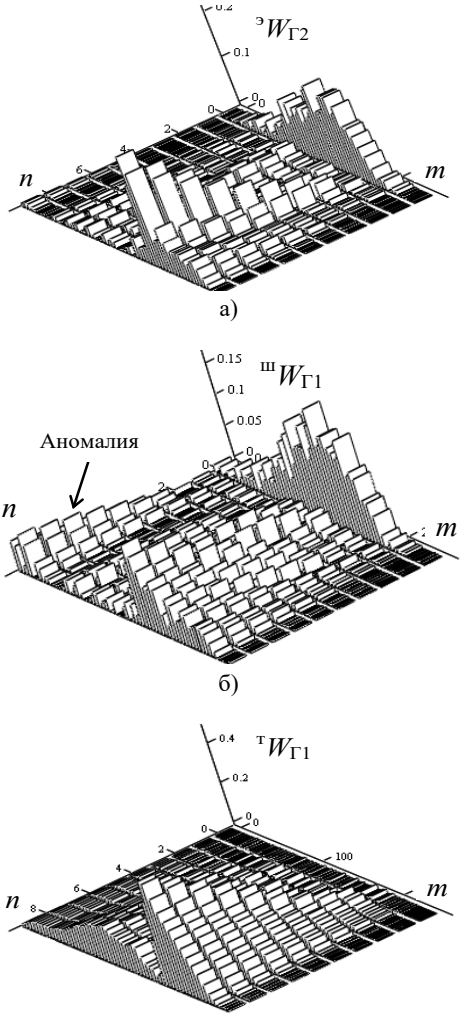


Рис. 12. а) матрицы эталонного изображения на основе вейвлета Гаусса 2-го порядка; б) матрицы изображения со шлюпкой на основе вейвлета Гаусса 2-го порядка; в) матрицы текущего изображения на основе вейвлета Гаусса 2-го порядка

Здесь ${}^{э}W_{Г2}$, ${}^{т}W_{Г2}$, ${}^{ш}W_{Г2}$ – ДВП эталонного изображения на основе Гаусса 2-го порядка.

Отметим, что даже визуальный анализ полученных изображений матриц ДВП позволяет сделать вывод о повышении контрастности признакового пространства.

Представленные на рисунках 9-11 матрицы ДВП для лучшей демонстрации построены с шагом сдвига равного четырем.

Так на матрицах ДВП, сформированных на основе вейвлетов Гаусса 1-го и 2-го порядков, для изображений шлюпки отчетливо просматриваются характеризующие ее вейвлет-коэффициенты (на рисунках 10(б) и 11(б) указаны стрелкой).

Выявленные особенности позволяют выдвинуть гипотезу о повышении контрастности признаковых пространств изображений в результате их формирования на основе вейвлет-преобразований.

По результатам визуального анализа было принято заключение, что наиболее рельефно вейвлет-коэффициенты, характеризующие наличие шлюпки, проявляются на матрице ДВП, построенной на основе вейвлета Гаусса 2-го порядка. Поэтому они были выбраны для построения векторов признаков.

На рисунке 13 представлены векторы признаков, сформированные на основе матриц ДВП путем их усреднения по всем масштабам на каждом временном сдвиге.

Отметим, что построение векторов признаков из матриц ДВП возможно, и по переменной n, и по переменной m. Выбор в пользу последней переменной m обусловлен результатами, полученными в [4].

Для оценки контрастности векторов признаков, сформированных на основе ДВП, использующего вейвлет Гаусса 2-го порядка, воспользуемся формулой (1).

Ниже представлены результаты взаимной контрастности векторов признаков между эталонным и текущим изображением, а также эталонным и изображением шлюпки.

$$ {}^{эш}d = \frac{1}{256}\sum_{m=0}^{255} | {}^{э}V_m - {}^{ш}V_m | = 0{,}15, $$

$$ {}^{эт}d = \frac{1}{256}\sum_{m=0}^{255} | {}^{э}V_m - {}^{т}V_m | = 0{,}019. $$

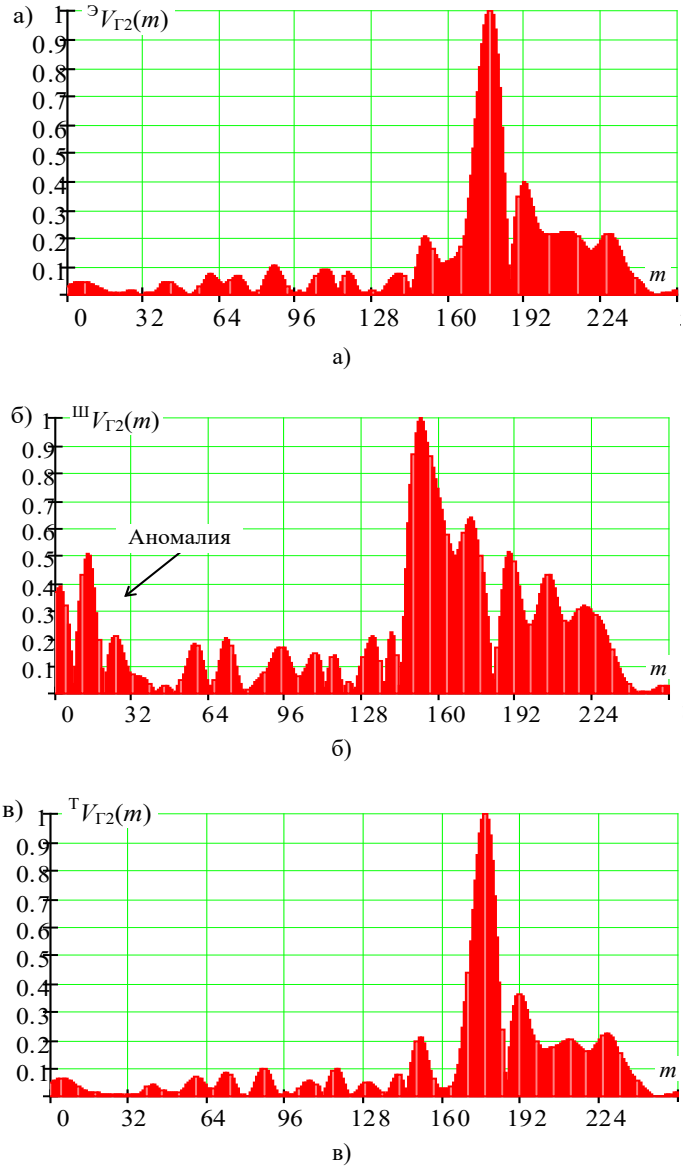


Рис. 13. а) вектор признаков эталонного изображения на основе вейвлета Гаусса 2-го порядка; б) вектор признаков изображения со шлюпкой на основе вейвлета Гаусса 2-го порядка; в) вектор признаков текущего изображения на основе вейвлета Гаусса 2-го порядка

Полученные результаты показывают, что для рассматриваемого примера различия в контрастности между парами разностей составляют почти 7,9 раз.

Следует отметить, что если в качестве эталона рассматривать текущее изображение, то различия увеличатся в 8,3 раза.

Для лучшей наглядности результатов исследования на рисунке 14 представлены функции различий между векторами признаков эталонного изображения и изображения со шлюпкой, полученные на основе гистограмм и ДВП с вейвлетом Гаусса 2-го порядка, рассчитанные по формуле

$$^{\text{эт}}d_{\text{ГГ/Г2}}(m) = |\ ^{\text{э}}V_m - {}^{\text{т}}V_m\ |, \tag{11}$$

где $^{\text{эт}}d_{\text{ГГ}}(m)$ – функция разности между векторами признаков нижний индекс эталонного изображения и изображения со шлюпкой, полученные на основе гистограмм;

$^{\text{эт}}d_{\text{Г2}}(m)$ – функция разности между векторами признаков нижний индекс эталонного изображения и изображения со шлюпкой, полученные на основе ДВП, использующего в качестве материнского вейвлета функцию Гаусса 2-го порядка.

Очевидно, что дисперсия вектора $^{\text{эт}}d_{\text{Г2}}(m)$ существенно превосходит аналогичный показатель вектора $^{\text{эт}}d_{\text{ГГ}}(m)$, что позволяет судить о повышении контрастности признакового пространства, сформированного на основе матриц ДВП.

Учитывая, что обнаружение, в данном случае спасательной шлюпки, будет происходить на фоне динамически изменяющейся морской поверхности, то такой процесс будет носить вероятностный характер.

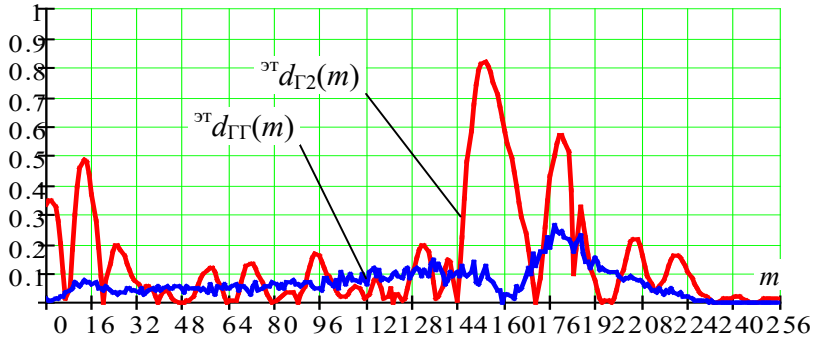


Рис. 14. Функции разности векторов признаков

Поскольку рассмотренный подход к обнаружению аномалий основан на статистической обработке параметров изображений, то открывается возможность оценки его эффективности, с позиций вероятностных показателей.

Подходы к вероятностной оценке распознавания объектов по результатам обработки изображений известны. Однако все они, как правило, ориентированы на определенные условия.

Так в [39], вероятность распознавания рассмотрена с позиций возможностей аппаратуры и условия наблюдения. Полученная в работе аналитическая зависимость носит эмпирический характер и содержит большое количество различных уточняющих коэффициентов и параметров. Однако в основу вероятностного выражения составляет величина отношения площади объекта к общей площади изображения.

В [40] получено аналитическое выражение вероятности правильного обнаружения с позиций обнаружения сигналов в шумах по критерию Байеса и Неймана–Пирсона, по результатам обработки отношения плотностей распределения вероятностей. В основе используемой аналитики авторы рассматривают отношение математического ожидания к дисперсии функции правдоподобия при условии наличия и отсутствия распознаваемого объекта на обрабатываемом изображении.

Очевидно, что такой подход правомерен, поскольку любая внесенная аномалия приводит к изменению статистических параметров изображения. При этом авторы справедливо отмечают, что такой подход применим в предположении «… гауссовости яркостей обрабатываемых изображений как в отсутствие сигнала, так и при его наличии …» [40]. Поэтому предлагают обнаруживать аномалии по результатам предварительной фильтрации общего фона. Но такое решение изначально предполагает высокую статичность фона, что сложно реализуемо в морских акваториях.

Вместе с тем анализ указанных подходов позволяет сделать вывод о том, что основными параметрами, определяющими качество распознавания, являются относительные размеры объекта распознавания к общему размеру кадра изображения и вариативность структуры окружающего идентифицируемый объект фона. Следовательно, эти параметры и должны определять величину искомой вероятности распознавания.

Поскольку в рассматриваемой ситуации в качестве обрабатываемых данных выступают векторы признаков, представленные на рисунке 12, то в основу подхода к распознаванию аномалий, приводящих к изменению их структуры, целесообразно

определить способы, используемые при обнаружении сигнала с неизвестной амплитудой.

Согласно [41], в таких способах величина условной вероятности правильного обнаружения $P_{\text{по}}$ определяется текущим значением отношения сигнал/шума h^2 на входе решающего устройства

$$P_{\text{по}} = \begin{cases} 0,5\left[1+\Phi(h-G_0)\right], & h \geq G_0; \\ 0,5\left[1-\Phi(G_0-h)\right], & h < G_0, \end{cases} \tag{12}$$

где G_0 – нормированный порог обнаружения; Ф(*) – функция Крампа.

$$\Phi(G_0) = \sqrt{\frac{2}{\pi}} \int_0^{Go} \exp\frac{-t^2}{2}\,\mathrm{d}t.$$

В терминах аналогий можно полагать, что в качестве шума, в рассматриваемом случае, будут выступать различия между исходным (эталонным) изображением и обрабатываемым.

В общем случае, задача вероятности обнаружения будет состоять в выявлении различий, вызванных наличием искомых аномалий на обрабатываемом изображении. Но поскольку в качестве фона изображений выступает водная поверхность, то даже в отсутствии аномалий, вызванных искомого объекта распознавания, эталонное и текущее изображения будут отличаться друг от друга.

Очевидно, что чем сильнее будут эти различия, тем сложнее будет выявить изменения, обусловленные наличием аномалий. В свою очередь изменения вектора признаков, вызванные наличием аномалий, будут зависеть от их размеров на обрабатываемом кадре изображения. Это связано с тем, что наличие аномалий приводит к амплитудным изменениям соответствующим им элементам векторов признаков. Поэтому чем крупнее объект (аномалия) в пределах обрабатываемого кадра, тем больше количество пикселов будут описывать его изображение. И, следовательно, соответствующие ему элементы вектора признаков получат большие амплитудные значения. Тем самым повысится контрастность вектора признаков.

Поэтому, с рассмотренных позиций, в формуле (12) величину h^2 предлагается рассматривать в виде

$$h^2 = B / \sigma, \tag{13}$$

где $B \in [0; 1]$ – относительные размеры объекта распознавания на кадре изображения, σ – величина среднего квадратического разброса параметров вектора признаков.

Ограничением предложенного подхода является условие, что различия между двумя любыми изображениями из рассматриваемого признакового пространства не должны превышать величины σ.

Значение порога G_0 в формуле (12) будет определять величину различий между векторами признаков эталонного и обрабатываемого изображений, вызванных наличием аномалий, превышение которых позволяет их идентифицировать (отнести к искомому объекту).

Тогда с учетом введенных допущений можно получить графики зависимости вероятности обнаружения в кадре изображения аномалии (искомого объекта) в заданном пространстве распознавания (рисунок 15).

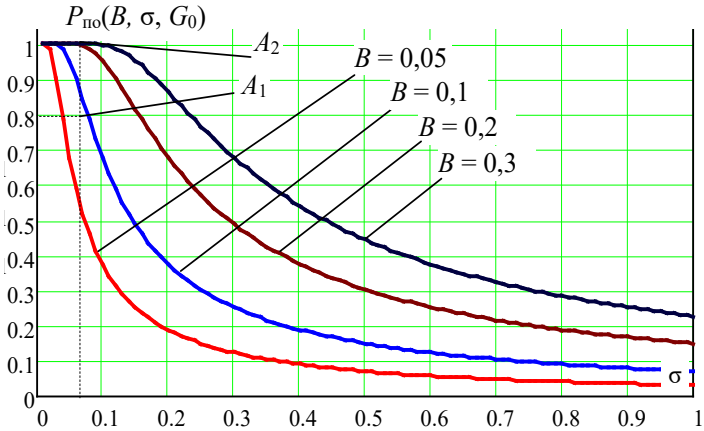


Рис. 15. Вероятность правильного обнаружения (распознавания) аномалий на кадрах изображений

Представленные на рисунке 15 зависимости получены при значении величины порога $G_0 = 0,01$.

Согласно полученным результатам, при величине $\sigma = 0,1$ обнаружение аномалий с вероятностью не менее 0,95 становится возможным, если их размеры будут занимать не менее 20% от общей площади обрабатываемого кадра изображения.

При этом снижение размера аномалии до 10% от общей площади кадра, приведет к снижению вероятности ее обнаружения до

уровня 0,7. А при размерах аномалий порядка 5%, вероятность их обнаружения снизится до значения 0,4.

В ходе моделирования использовались 200 различных изображений, для которых векторы признаков строились на основе гистограмм распределения яркости.

Так, на рисунке 6 представлено текущее изображение (из числа используемых для моделирования), которое имеет наибольшие различия с эталонным изображением (рисунок 3). В рассматриваемом случае величина средних квадратических различий достигает σ = 0,07.

При этом размер шлюпки занимает порядка 7% от общей площади кадра изображения, поэтому вероятность правильного обнаружения достигает величины $P_{\text{по}} = 0{,}8$ (точка A_1 на рисунке 15).

Заметим, что в представленном варианте выражение для расчета вероятности правильного обнаружения не учитывает дополнительную контрастность, которая обеспечивается при переходе к формированию векторов признаков на основе реализации процедур ДВП.

Для учета контрастности, обеспечиваемой процедурами ДВП, предлагается использовать коэффициент повышения контрастности $K_{\text{пк}}$, который можно рассчитать, применительно к проведенному исследованию, как

$$K_{\text{пк}} = \frac{{}^{\text{эш}}d_{\Gamma 2} \big/ {}^{\text{эт}}d_{\Gamma 2}}{{}^{\text{эш}}d_{\Gamma\Gamma} \big/ {}^{\text{эт}}d_{\Gamma\Gamma}}, \tag{14}$$

где ${}^{\text{эш}}d_{\Gamma\Gamma}/{}^{\text{эт}}d_{\Gamma\Gamma}$ – отношение векторов разности, сформированных на основе гистограмм распределения яркости; ${}^{\text{эш}}d_{\Gamma 2}/{}^{\text{эт}}d_{\Gamma 2}$ – отношение векторов разности, сформированных на основе ДВП в базисе вейвлетов Гаусса 2-го порядка.

Для рассмотренных условий:

${}^{\text{эш}}d_{\Gamma\Gamma}/{}^{\text{эт}}d_{\Gamma\Gamma} = 2{,}3$;

${}^{\text{эш}}d_{\Gamma 2}/{}^{\text{эт}}d_{\Gamma 2} = 7{,}9$.

Следовательно, коэффициент повышения контрастности будет равен $K_{\text{пк}} = 3{,}4$.

Поскольку повышение контрастности эквивалентно увеличению размера распознаваемого объекта, то результирующее выражение для оценки вероятности правильного обнаружения аномалий с учетом коэффициента повышения контрастности можно записать как

$$P_{\text{по}} = \begin{cases} 0,5\left[1+\Phi\left(\dfrac{BK_{\text{пк}}}{\sigma}-G_0\right)\right], & \dfrac{BK_{\text{пк}}}{\sigma} \geq G_0; \\ 0,5\left[1-\Phi\left(G_0-\dfrac{BK_{\text{пк}}}{\sigma}\right)\right], & \dfrac{BK_{\text{пк}}}{\sigma} < G_0. \end{cases} \quad (15)$$

С учетом выражения (15), т.е. при формировании векторов признаков на основе ДВП, вероятность правильного обнаружения аномалий достигает величины $P_{\text{по}} = 0,99$ (точка A_2 на рисунке 14).

Отметим, что по аналогии [42], используемой для обоснования вероятности правильного обнаружения, можно получить и величину вероятности ложной тревоги

$$P_{\text{лт}} = 0,5\left[1-\Phi\left(\frac{G_0}{\sqrt{2}B/\sigma}\right)\right]. \quad (16)$$

Полученные результаты указывают на достижение поставленной цели.

6. Заключение. Полученные результаты указывают на правомерность выбранного подхода к повышению контрастности векторов признаков, позволивших повысить достоверность выявления аномалий на изображениях при формировании их векторов признаков в базисах вейвлетов. Дальнейшие исследования авторы связывают с более детальной проработкой метода вероятностной оценки результатов распознавания.

Литература

1. Киджи Д.С., Ольховик Е.О. Методика районирования акватории Северного морского пути при проведении аварийноспасательных работ и ликвидации разливов нефти // Вестник государственного университета морского и речного флота им. адмирала С.О. Макарова. 2023. Т. 15. № 6. С. 1030–1040. DOI: 10.21821/2309-5180-2023-15-6-1030-1040.
2. Абросимов В.К., Матвеева Ю.Н. Формирование синтетических данных для машинного обучения распознаванию подводных объектов // Робототехника и техническая кибернетика. 2023. Т. 11. № 4. С. 256–266. DOI: 10.31776/RTCJ.11402.
3. Кузьмин О.В., Лавлинская А.А., Тараканов Б.А., Федоренко М.А., Былков Е.Г., Харитонов И.А. Проектирование БПЛА для мониторинга лесных массивов и водных пространств при помощи Autodesk Fusion 360 // Информационные технологии, их приложения и информационное образование: Материалы II Международной научной конференции (г. Улан-Удэ – Гусиноозерск, 20–22 августа 2021 г.) Улан-Удэ: Бурятский государственный университет имени Доржи Банзарова, 2021. С. 168–177.

4. Дворников С.В., Дворников С.С., Васильева Д.В. Автоматизация процедур обнаружения нефтяных разливов на водной поверхности // Информация и космос. 2024. № 1. С. 126–132.
5. Бударова В.А., Воронина Е.А., Дубровский А.В., Кустышева И.Н., Малыгина О.И., Мартынова Н.Г., Торсунова О.Ф. Нормативно-правовые особенности установления водоохранных зон и прибрежных защитных полос (на примере территории Новосибирской области) // Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий). 2020. Т. 25. № 1. С. 222–238. DOI: 10.33764/2411-1759-2020-25-1-222-238.
6. Martyshov M.I., Nikitenko D.A. Preprocessing of system monitoring data for workload analysis of HPC systems // Numerical Methods and Programming. 2021. vol. 22. no. 3. pp. 230–238. DOI 10.26089/NumMet.v22r314.
7. Родионов В.В., Ищук И.Н., Громов Ю.Ю. Применение технологий искусственного интеллекта в задачах обработки данных дистанционного мониторинга поверхности земли // Автоматизация в промышленности. 2024. № 1. С. 26–28. DOI: 10.25728/avtprom. 2024.01.06.
8. Горбунова Э.М., Беседина А.Н., Кабыченко Н.В., Батухтин И.В., Петухова С.М. Прецизионный гидрогеологический мониторинг в техногенно-нарушенных условиях: организация, проведение и обработка экспериментальных данных // Сейсмические приборы. 2021. Т. 57. № 2. С. 62–80. DOI: 10.21455/si.2021.2-4.
9. Чан Х.Н., Подстригаев А.С, Нгуен Ч.Н., Иконенко Д.А. Оценка достоверности распознавания сигналов для алгоритма распознавания с детектированием на двух промежуточных частотах // Успехи современной радиоэлектроники. 2023. Т. 77. № 10. С. 70–79. DOI: 10.18127/j20700784-202310-07.
10. Васильева Д.В., Якушенко С.А., Дворников С.В., Погорелов А.А., Дворников С.В. Обнаружение морских дронов в оптическом диапазоне // Морской вестник. 2023. № 4(88). С. 90–92.
11. Ван Л., Петросян О.Г. Распознавание лиц на основе классификации вейвлет признаков путём вейвлет нейронных сетей // Информатизация образования и науки. 2018. № 4(40). С. 129–139.
12. Гочаков А.В. Задача распараллеливания вейвлет-преобразования для фильтрации зашумленных видеоизображений // Труды Новосибирского государственного архитектурно-строительного университета (Сибстрин). 2011. Т. 14. № 2(51). С. 85–89.
13. Dvornikov S., Ustinov A., Okov I. Statistical Arithmetic Coding Algorithm Adaptive to Correlation Properties of Wavelet Transform Coefficients // Proceedings of Telecommunication Universities. 2022. vol. 8. no. 3. pp. 6–12. DOI: 10.31854/1813-324X-2022-8-3-6-12.
14. Дубина В.А., Катин И.О., Боброва М.А., Плотников В.В. Кораблекрушение на границе морского заповедника. Результаты спутникового мониторинга // Современные проблемы дистанционного зондирования Земли из космоса. 2020. Т. 17. № 1. С. 267–270. DOI: 10.21046/2070-7401-2020-17-1-267-270.
15. Родионов А.Г., Ефимов В.В., Тварин Ю.Г. Искусственный интеллект в судовождении - игра в имитацию? (По материалам зарубежных и открытых источников) // Морской вестник. 2022. № 4(84). С. 95–102.
16. Сенцов А.А., Ненашев В.А., Иванов С.А., Турнецкая Е.Л. Совмещение сформированных радиолокационных изображений с цифровой картой местности в бортовых системах оперативного мониторинга земной поверхности // Труды МАИ. 2021. № 117. DOI: 10.34759/trd-2021-117-08.
17. Гибадуллин Р.Ф., Смирнов И.Н., Хевронин Н.В., Никитин А.В., Перухин М.Ю. Разработка аппаратно-программного модуля обнаружения объектов для

встраиваемых систем // Вестник Технологического университета. 2018. Т. 21. № 6. С. 118–122.
18. Janati M., Kolahdoozan M., Imanian H. Artificial Neural Network Modeling for the Management of Oil Slick Transport in the Marine Environments. Pollution. 2020. vol. 6. no. 2. pp. 399–415.
19. Jiao Z., Jia G., Cai Y. A new approach to oil spill detection that combines deep learning with unmannedaerial vehicles. Computers & Industrial Engineering. 2019. vol. 135. pp. 1300–1311.
20. Stockman G., Shapiro L.G. Computer Vision. 1st ed. Prentice Hall PTR: Upper Saddle River, NJ, USA, 2001. 617 p.
21. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016. pp. 770–778.
22. Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017.
23. Фукунага К. Введение в статистическую теорию распознавания образов / Пер. с англ. Москва: Наука, 1979. 368 с.
24. Буланов В.А., Волосатова Т.М. Программный комплекс предварительной обработки изображений для обнаружения и распознавания изображений // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. 2014. № 4. С. 321–338. DOI: 10.7463/0414.0707888.
25. Abbasov I.B. Image Recognition in Agriculture and Landscape Protection // International Journal of Science and Research. 2020. vol. 9. no. 12. pp. 757–763. DOI: 10.21275/SR201212144831.
26. Dvornikov S.S., Zheglov K.D., Dvornikov S.V. SSB signals with controlled pilot level // T-Comm. 2023. vol. 17. no. 3. pp. 41–47. DOI: 10.36724/2072-8735-2023-17-3-41-47.
27. Zimek A., Schubert E. Outlier Detection // Encyclopedia of Database Systems. 2017. DOI: 10.1007/978-1-4899-7993-3_80719-1.
28. Васильева Д.В., Дворников С.С., Толстуха Ю.Е., Обрезков П.С., Дворников С.В. Формирование векторов признаков для систем видеонаблюдения // Вопросы радиоэлектроники. Серия: Техника телевидения. 2023. № 4. С. 62–68.
29. Казарян М.Л., Рихтер А.А., Шахраманьян М.А. Символические матрицы и их применение при решении систем многочленных уравнений и обработке изображений // Информация и космос. 2021. № 3(1). С. 86–95.
30. Дворников С.В., Балыков А.А. Предложения по управлению скоростью передачи и помехоустойчивостью сигналов с перестановочной частотной модуляцией // T-Comm: Телекоммуникации и транспорт. 2020. Т. 14. № 6. С. 20–26. DOI: 10.36724/2072-8735-2020-14-6-20-26.
31. Макаренко А.А. Вариант применения цифровой обработки изображений для распознавания текстур на оптикоэлектронном изображении // Радиопромышленность. 2021. Т. 31. № 2. С. 15–21. DOI: 10.21778/2413-9599-2021-31-2-15-21.
32. The First Half-Tones. Library and Archives of Canada. 2009.
33. Умбиталиев А.А., Дворников С.В., Оков И.Н., Устинов А.А. Способ сжатия графических файлов методами вейвлет-преобразований // Вопросы радиоэлектроники. Серия: Техника телевидения. 2015. № 3. С. 100–106.
34. Горбачев В.Н., Казаков А.Я., Савельева М.Ю. Вейвлет-преобразование полутонового изображения в конечном поле // Оптический журнал. 2021. Т. 88. № 2. С. 40–49. DOI: 10.17586/1023-5086-2021-88-02-40-49.

35. Дейцева А.Г. Построение базисных вейвлетов и фреймов с помощью финитных функций. Соболевский вейвлет // Вестник БГУ. Серия 1, Физика. Математика. Информатика. 2006. № 2. С. 75–79.
36. Велигоша А.В., Малышко Н.Н., Струков Р.И. Применение искусственных нейронных сетей для снижения избыточности дискретного вейвлет-преобразования // Теория и техника радиосвязи. 2020. № 2. С. 5–12.
37. Сай С.В., Зинкевич А.В., Фомина Е.С. Сравнение дискретного косинус и вейвлет-преобразований в системах сжатия RAW-изображений // Компьютерная оптика. 2022. Т. 46. № 6. С. 929–938. DOI: 10.18287/2412-6179-CO-1094.
38. Лаврентьева А.С., Новиков В.А., Семенова М.Ю., Фидельман В.Р. Применение дискретного вейвлет-преобразования для определения символьной скорости коротких выборок фазоманипулированного сигнала // Известия высших учебных заведений. Поволжский регион. Технические науки. 2016. № 1(37). С. 92–102.
39. Сельвесюк Н.И., Веселов Ю.Г., Гайденков А.В., Островский А.С. Оценка характеристик обнаружения и распознавания объектов на изображении от специальных оптико-электронных систем наблюдения летного поля // Труды МАИ. 2018. № 103.
40. Андриянов Н.А., Дементьев В.Е., Ташлинский А.Г. Обнаружение объектов на изображении: от критериев Байеса и Неймана–Пирсона к детекторам на базе нейронных сетей EfficientDet // Компьютерная оптика. 2022. Т. 46. № 1. С. 139–159. DOI: 10.18287/2412-6179-CO-922.
41. Дворников С.В. Метод обнаружения сигналов диапазона ВЧ на основе двухэтапного алгоритма принятия решения // Научное приборостроение. 2005. Т. 15. № 3. С. 114–119.

Дворников Сергей Викторович — д-р техн. наук, профессор кафедры, кафедра радиотехнических и оптоэлектронных комплексов, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения» (ГУАП); кафедра радиосвязи, Военная академия связи им. Маршала Советского Союза С.М. Буденного. Область научных интересов: построение помехозащищенных систем радиосвязи, формирование и обработка сигналов сложных структур. Число научных публикаций — 448. practicdsv@yandex.ru; Тихорецкий проспект, 3, 194064, Санкт-Петербург, Россия; р.т.: +7(812)247-9400.

Васильева Дина Владимировна — аспирант, старший преподаватель кафедры, кафедры радиотехнических систем института радиотехники и инфокоммуникационных технологий, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет аэрокосмического приборостроения» (ГУАП). Область научных интересов: цифровая обработка сигналов. Число научных публикаций — 15. dolli.dina@mail.ru; улица Большая Морская, 67А, 190000, Санкт-Петербург, Россия; р.т.: +7(931)385-9181.

DOI 10.15622/ia.23.6.5

S. DVORNIKOV, D. VASILIEVA

INCREASE OF RELIABILITY OF ANOMALIES DETECTION ON IMAGES AT FORMATION OF THEIR FEATURE VECTORS IN WAVELET BASES

Dvornikov S., Vasilieva D. **Increase of Reliability of Anomalies Detection on Images at Formation of Their Feature Vectors in Wavelet Bases.**

Abstract. The method of detection of life rafts and lifeboats in the water area of seas and oceans after shipwrecks based on the recognition of anomalies in the processed images, which increases the probability of recognition of monitoring objects, is proposed. The approach to solving such a problem is substantiated. The formulation of the problem of object recognition from the perspective of binary classification in the detection of anomalies is presented. The analytical expression for the decision-making algorithm is obtained. The possibility of formalization of image matrices in the form of histograms of color (brightness) intensity distributions is considered. The contrast of the feature space on their basis is estimated. It is suggested that the contrast of feature spaces be increased due to the secondary processing of histograms of distributions in the basis of multiple-scale wavelet decomposition. The possibility of realization of wavelet transformations on the basis of Haar functions and Gauss wavelets of the 1st and 2nd orders is considered. The mechanism of formation of secondary feature vectors from three-dimensional wavelet transforms by averaging their coefficients along the time shift axis is substantiated. It is shown that at the same dimensionality of histograms of brightness distribution with newly formed feature vectors, the latter provide higher contrast of feature spaces. It is recommended to use a Gaussian wavelet of the 2nd order for the formalization of images in jpeg format, which provides, other things being equal, a greater magnitude of differences for images containing anomalies. An approach to probabilistic evaluation of the algorithm for automatic image recognition is developed. The analytical expression is obtained and its constituent elements are justified. Graphical dependences of the probability of correct detection (recognition) of anomalies, depending on the size in relation to the total area of the frame and the dispersion of the underlying background are given. The results of the experiment on image recognition with a lifeboat in the ocean water area are presented. The directions of further research are defined.

Keywords: image recognition, feature vectors, wavelet transform, contrast of feature spaces.

References

1. Kiji D.S., Olkhovik E.O. [Methodology of zoning the water area of the Northern Sea Route during emergency rescue operations and oil spill response]. Vestnik gosudarstvennogo universiteta morskogo i rechnogo flota im. admirala S.O. Makarova – Bulletin of the State University of Sea and River Fleet named after Admiral S.O. Makarov. 2023. vol. 15. no. 6. pp. 1030–1040. DOI: 10.21821/2309-5180-2023-15-6-1030-1040. (In Russ.).
2. Abrosimov V.K., Matveeva Yu.N. [Formation of synthetic data for machine learning to recognize underwater objects]. Robototekhnika i tekhnicheskaya kibernetika – Robotics and Technical Cybernetics. 2023. vol. 11. no. 4. pp. 256–266. DOI: 10.31776/RTCJ.11402. (In Russ.).
3. Kuzmin O.V., Lavlinskaya A.A., Tarakanov B.A., Fedorenko M.A., Bylkov E.G., Haritonov I.A. [Designing UAVs for monitoring forests and water spaces using Autodesk Fusion 360] Informacionnye tekhnologii, ih prilozheniya i informacionnoe

obrazovanie: Materialy II Mezhdunarodnoj nauchnoj konferencii [Information technologies, their applications and information education: Proceedings of the II International Scientific Conference]. Ulan-Ude: Dorzhi Banzarov Buryat State University. 2021. pp. 168–177. (In Russ.).

4. Dvornikov S.V., Dvornikov S.S., Vasilieva D.V. [Automation of the oil spill detection procedures on the water surface]. Informaciya i kosmos – Information and Space. 2024. no. 1. pp. 126–132. (In Russ.).
5. Budarova V.A., Voronina E.A., Dubrovsky A.V., Kustysheva I.N., Malygina O.I., Martynova N.G., Torsunova O.F. [Normative and legal features of establishing water protection zones and coastal protective strips (by the example of the territory of the Novosibirsk region)]. Vestnik SGUGiT (Sibirskogo gosudarstvennogo universiteta geosistem i tekhnologij) – Bulletin of SGUGiT (Siberian State University of Geosystems and Technologies). 2020. vol. 25. no. 1. pp. 222–238. DOI: 10.33764/2411-1759-2020-25-1-222-238. (In Russ.).
6. Martyshov M.I., Nikitenko D.A. Preprocessing of system monitoring data for workload analysis of HPC systems. Numerical Methods and Programming. 2021. vol. 22. no. 3. pp. 230–238. DOI 10.26089/NumMet.v22r314.
7. Rodionov V.V., Ishchuk I.N., Gromov Yu.Yu. [Application of artificial intelligence technologies in the tasks of processing the data of remote monitoring of the earth surface]. Avtomatizaciya v promyshlennosti – utomation in industry. 2024. no. 1. pp. 26–28. DOI: 10.25728/avtprom. 2024.01.06. (In Russ.).
8. Gorbunova E.M., Besedina A.N., Kabychenko N.B., Batuhtin I.V., Petuhova S.M. [Precision hydrogeological monitoring in technogenically disturbed conditions: organization, conduct and processing of experimental data]. Sejsmicheskie pribory – Seismic Devices. 2021. vol. 57. no. 2. pp. 62–80. DOI: 10.21455/si.2021.2-4. (In Russ.).
9. Chan H.N., Podstrigaev A.S., Nguyen C.N., Ikonenko D.A. [Estimation of signal recognition reliability for the recognition algorithm with detection at two intermediate frequencies]. Uspekhi sovremennoi radioelectronics – Advances in modern radio electronics. 2023. vol. 77. no. 10. pp. 70–79. DOI: 10.18127/j20700784-202310-07. (In Russ.).
10. Vasilieva D.V., Yakushenko S.A., Dvornikov S.B., Pogorelov A.A., Dvornikov S.V. [Detection of marine drones in the optical range]. Morskoj vestnik – Marine Bulletin. 2023. no. 4(88). pp. 90–92. (In Russ.).
11. Wang L., Petrosyan O.G. [Face recognition based on wavelet feature classification by wavelet neural networks]. Informatizaciya obrazovaniya i nauki – Informatization of education and science. 2018. no. 4(40). pp. 129–139. (In Russ.).
12. Gochakov A.V. [Parallelization problem wavelet transform for filtering of noisy video images] Trudy Novosibirskogo gosudarstvennogo arhitekturno-stroitel'nogo universiteta (Sibstrin) [Proceedings of Novosibirsk State University of Architecture and Civil Engineering (Sibstrin)]. 2011. vol. 14. no. 2(51). pp. 85–89. (In Russ.).
13. Dvornikov S., Ustinov A., Okov I. Statistical Arithmetic Coding Algorithm Adaptive to Correlation Properties of Wavelet Transform Coefficients. Proceedings of Telecommunication Universities. 2022. vol. 8. no. 3. pp. 6–12. DOI: 10.31854/1813-324X-2022-8-3-6-12.
14. Дубина В.А., Катин И.О., Боброва М.А., Плотников В.В. Кораблекрушение на границе морского заповедника. Результаты спутникового мониторинга. Современные проблемы дистанционного зондирования Земли из космоса. 2020. Т. 17. № 1. С. 267–270. DOI: 10.21046/2070-7401-2020-17-1-267-270.
15. Rodionov A.G., Efimov V.V., Tvarin Yu.G. [Artificial intelligence in ship navigation – a game of imitation? (On the materials of foreign and open sources)]. Morskoj vestnik – Marine Bulletin. 2022. no. 4(84). pp. 95–102. (In Russ.).

16. Sentsov A.A., Nenashev V.A., Ivanov S.A., Tournetskaya E.L. [Alignment of the generated radar images with the digital terrain map in the onboard systems of the operational monitoring of the Earth surface] Trudy MAI [Proceedings of MAI]. 2021. no. 117. DOI: 10.34759/trd-2021-117-08. (In Russ.).
17. Gibadullin R.F., Smirnov I.N., Khevronin N.B., Nikitin A.V., Peruhin M.Yu. [Development of hardware and software object detection module for embedded systems]. Vestnik Tekhnologicheskogo universiteta – Vestnik of Technological University. 2018. vol. 21. no. 6. pp. 118–122. (In Russ.).
18. Janati M., Kolahdoozan M., Imanian H. Artificial Neural Network Modeling for the Management of Oil Slick Transport in the Marine Environments. Pollution. 2020. vol. 6. no. 2. pp. 399–415.
19. Jiao Z., Jia G., Cai Y. A new approach to oil spill detection that combines deep learning with unmannedaerial vehicles. Computers & Industrial Engineering. 2019. vol. 135. pp. 1300–1311.
20. Stockman G., Shapiro L.G. Computer Vision. 1st ed. Prentice Hall PTR: Upper Saddle River, NJ, USA, 2001. 617 p.
21. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016. pp. 770–778.
22. Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017.
23. Fukunaga K. Vvedenie v statisticheskuyu teoriyu raspoznavaniya obrazov [Introduction to the statistical theory of pattern recognition]. Moscow: Nauka, 1979. 367 p. (In Russ.).
24. Bulanov V.A., Volosatova T.M. [Program complex of preliminary image processing for image detection and recognition]. Nauka i obrazovanie: nauchnoe izdanie MGTU im. N.E. Baumana – Science and Education: scientific edition of Bauman Moscow State Technical University. 2014. no. 4. pp. 321–338. DOI: 10.7463/0414.0707888. (In Russ.).
25. Abbasov I.B. Image Recognition in Agriculture and Landscape Protection. International Journal of Science and Research. 2020. vol. 9. no. 12. pp. 757–763. DOI: 10.21275/SR201212144831.
26. Dvornikov S.S., Zheglov K.D., Dvornikov S.V. SSB signals with controlled pilot level. T-Comm. 2023. vol. 17. no. 3. pp. 41–47. DOI: 10.36724/2072-8735-2023-17-3-41-47.
27. Zimek A., Schubert E. Outlier Detection. Encyclopedia of Database Systems. 2017. DOI: 10.1007/978-1-4899-7993-3_80719-1.
28. Vasilieva D.V., Dvornikov S.S., Tolstukha Y.E., Obrezkov P.S., Dvornikov S.V. [Formation of feature vectors for video surveillance systems]. Voprosy radioelektroniki. Seriya: Tekhnika televideniya – Questions of Radio Electronics. Series: Television Technology. 2023. no. 4. pp. 62–68. (In Russ.).
29. Kazarian M.L., Richter A.A., Shahramanian M.A. [Symbolic matrices and their application in solving systems of polynomial equations and image processing]. Informaciya i kosmos – Information and Space. 2021. no. 3. pp. 86–95. (In Russ.).
30. Dvornikov S.V., Balykov A.A. [Proposals for controlling the transmission rate and noise immunity of signals with permutation frequency modulation]. T-Comm: Telekommunikacii i transport – T-Comm: Telecommunications and Transportation. 2020. vol. 14. no. 6. pp. 20–26. DOI: 10.36724/2072-8735-2020-14-6-20-26. (In Russ.).
31. Makarenko A.A. [Variant of application of digital image processing for texture recognition on the optoelectronic image]. Radiopromyshlennost' – Radio Industry. 2021. vol. 31. no. 2. pp. 15–21. DOI: 10.21778/2413-9599-2021-31-2-15-21. (In Russ.).

32. The First Half-Tones. Library and Archives of Canada. 2009.
33. Umbitaliev A.A., Dvornikov S.V., Okov I.N., Ustinov A.A. lMethod of compression of graphic files by methods of wavelet transformations]. Voprosy radioelektroniki. Seriya: Tekhnika televideniya – Questions of Radio Electronics. Series: Television Technology. 2015. no. 3. pp. 100–106. (In Russ.).
34. Gorbachev V.N., Kazakov A.Y., Savelyeva M.Yu. [Wavelet transformation of a halftone image in a finite field]. Opticheskij zhurnal – Optical Journal. 2021. vol. 88. no. 2. pp. 40–49. DOI: 10.17586/1023-5086-2021-88-02-40-49. (In Russ.).
35. Дейцева А.Г. [Deitseva A.G. Construction of basis waveletsand frames with the help of finite functions. Sobolev wavelet]. Vestnik BGU. Seriya 1, Fizika. Matematika. Informatika – BSU Bulletin. Series 1, Physics. Mathematics. Computer Science. 2006. no. 2. pp. 75–79. (In Russ.).
36. Veligosha A.V., Malyshko N.N., Strukov R.I. [Application of artificial neural networks to reduce the redundancy of discrete wavelet transform] Teoriya i tekhnika radiosvyazi – Theory and technique of radio communication. 2020. no. 2. pp. 5–12. (In Russ.).
37. Say S.V., Zinkevich A.V., Fomina E.C. [Comparison of discrete cosine and wavelet transforms in RAW-image compression systems]. Komp'yuternaya optika – Computer Optics. 2022. vol. 46. no. 6. pp. 929–938. DOI: 10.18287/2412-6179-CO-1094. (In Russ.).
38. Lavrentieva A.S., Novikov V.A., Semenova M.Yu., Fidelman V.R. [Application of discrete wavelet transformation for the Determination of the symbolic rate of short samples of a phase-manipulated signal]. Izvestiya vysshih uchebnyh zavedenij. Povolzhskij region. Tekhnicheskie nauki – News of higher educational institutions. Volga region. Technical sciences. 2016. no. 1(37). pp. 92–102. (In Russ.).
39. Selvesyuk N.I., Veselov Y.G., Gaidenkov A.V., Ostrovsky A.S. [Evaluation of the characteristics of detection and recognition of objects in the image from special optical-electronic systems of airfield observation] Trudy MAI [Proceedings of MAI]. 2018. no. 103. (In Russ.).
40. Andriyanov N.A., Dementiev V.E., Tashlinsky A.G. [Object detection on the image: from the Bayes and Neyman-Pearson criteria to the detectors based on the EfficientDet neural networks]. // Komp'yuternaya optika – Computer Optics. vol. 46. no. 1. pp. 139–159. DOI: 10.18287/2412-6179-CO-922 (In Russ.).
41. Dvornikov S.V. [Method of detection of RF signals on the basis of two-stage algorithm of decision making]. Nauchnoe instrumentostroenie – Scientific instrumentation. 2005. vol. 15. no. 3. pp. 114–119. (In Russ.).

Dvornikov Sergey — Ph.D., Dr.Sci., Professor of the department, Department of radio engineering and optoelectronic complexes, Federal State Autonomous Educational Institution of Higher Education «St. Petersburg State University of Aerospace Instrumentation» (SUAI); Department of radio communications, Military Academy of Communications named after Marshal of the Soviet Union S.M. Budyonny. Research interests: construction of noise-protected radio communication systems, formation and processing of signals of complex structures. The number of publications — 448. practicdsv@yandex.ru; 3, Tikhoretsky Av., 194064, St. Petersburg, Russia; office phone: +7(812)247-9400.

Vasilieva Dina — Postgraduate student, senior lecturer of the department, Department of radio engineering systems, institute of radio engineering and info-communication technologies, Federal State Autonomous Educational Institution of Higher Education «St. Petersburg State University of Aerospace Instrumentation» (SUAI). Research interests: digital signal processing. The number of publications — 15. dolli.dina@mail.ru; 67A, Bolshaya Morskaya St., 190000, St. Petersburg, Russia; office phone: +7(931)385-9181.

УДК 004.89 **DOI** 10.15622/ia.23.6.6

В.Ф. Столярова , Т.В. Тулупьева , А.А. Вяткин

ПОДХОДЫ К ОЦЕНИВАНИЮ КУМУЛЯТИВНЫХ ХАРАКТЕРИСТИК ПОВЕДЕНИЯ В ГРУППАХ РАЗНОРОДНЫХ ИНДИВИДОВ: ТОЧНОСТЬ И ПРИМЕНИМОСТЬ В УСЛОВИЯХ ОГРАНИЧЕННЫХ НАБЛЮДЕНИЙ

Столярова В.Ф., Тулупьева Т.В., Вяткин А.А. **Подходы к оцениванию кумулятивных характеристик поведения в группах разнородных индивидов: точность и применимость в условиях ограниченных наблюдений.**

Аннотация. В ряде социоориентированных областей знаний возникает задача оценки кумулятивных характеристик поведения индивидов, таких как частота, которые реализуются в группах индивидов, причем поступающие данные сопряжены с неопределенностью. Нередки ситуации ограниченных данных, когда для небольшого числа наблюдаемых объектов известны лишь несколько эпизодов. Существуют несколько подходов, позволяющих строить оценки искомой кумулятивной характеристики в условиях ограниченных ресурсов: классический подход регрессии Кокса, оценка параметра копулы, апостериорный вывод в байесовских сетях доверия, классических и гибридных, однако до сих пор не были проанализированы возможности применимости имеющихся методов. Целью работы является анализ особенностей применения существующих методов косвенного оценивания интенсивности рискованного поведения индивидов на основе ограниченных данных об эпизодах для определения рекомендаций по их применению: определение точности оценок, получаемых с помощью перечисленных подходов, на основе расстояния Канторовича-Рубинштейна от истинного распределения искомой частоты, а также выявление требований к данным, которые предъявляются для построения оценок. Было показано, что подход на основе копул дает самые точные оценки и обладает наименьшими требованиями к количеству наблюдаемых объектов, однако не может учитывать внешние факторы, которые могут оказывать влияние на реализацию эпизодов поведения. Среди моделей, позволяющих учитывать ковариванты процесса, наибольшей точностью обладают оценки, опирающиеся на апостериорный вывод в гибридных байесовских сетях доверия. Полученные результаты являются новыми, они нацелены на применение в автоматизированных системах обработки информации о поведении индивидов. Практической значимостью обладают рекомендации по применению имеющихся подходов в зависимости от имеющихся данных.

Ключевые слова: последние эпизоды, рискованное поведение, байесовские сети доверия, гибридные байесовские сети доверия, копула.

1. Введение. В ряде областей знаний возникают задачи анализа последовательностей эпизодов некоторого поведения, которые реализуются в группах различающихся индивидов. Основным источником таких задач выступают социоориентированные области знаний, где поведение может быть сопряжено с риском. Для оценки показателей такого риска требуется выявлять характеристики, отражающие последовательности риск-ассоциированных эпизодов. Эти характеристики также позволяют сравнивать группы индивидов

между собой в плане поведения или с некоторым установленным значением. Интенсивность является одной из таких характеристик и широко используется при анализе риска в сфере эпидемиологии при мониторинге неизлечимых заболеваний [1, 2] и угрозообразующего поведения [3, 4], а также при анализе затрат в системе здравоохранения [5]. В области кибербезопасности возникают схожие задачи, связаные с анализом цифровых следов пользователей онлайн социальных сетей, установлением взаимосвязи между активностью индивидов в онлайн среде и особенностями личности [6 – 8], анализом распространения социоинженерных атакующих воздействий [9].

Однако сбор информации о риск-ассоциированном поведении может быть сопряжен со значительной неопределенностью, так как некоторые виды поведения труднодоступны для наблюдения и могут характеризоваться лишь в рамках самоотчетов самих индивидов. Для косвенной оценки искомых кумулятивных характеристик в условиях значительной неопределенности поступающей информации можно прибегать к математическому моделированию. В этом случае накладываются некоторые предположения, которые лежат в основе формализации процесса обработки неточной информации и разработки подходов к оцениванию интенсивности поведения. Естественной математической моделью для последовательностей риск-ассоциированных эпизодов являются точечные случайные процессы и гамма-пуассоновский процесс в частности [10, 11], возникает гамма-пуассоновская модель поведения. В этом случае основной характеристикой, отражающей паттерн реализации эпизодов, выступает *функция интенсивности* и для численного описания паттернов необходимо определить ее вид. В работе [12] было показано, что для гамма-пуассоновского процесса, функция интенсивности задается смешивающим гамма-распределением. Существуют несколько подходов к ее восстановлению по данным с неопределенностью: параметрический подход оценки параметра копулы и приближенные подходы классических и гибридных байесовских сетей доверия (БСД). Также подгонка регрессии Кокса позволяет получить оценку искомого параметра. Актуальной является задача анализа свойств этих подходов и выявления особенностей их применения для достижения наибольшей эффективности вычисления интенсивности в практических приложениях.

Целью работы является анализ особенностей применения существующих методов косвенного оценивания интенсивности рискованного поведения индивидов на основе ограниченных данных об эпизодах для определения рекомендаций по их применению.

Рассматриваемые методы имеют вероятностный характер, и потому исследование опирается на методы теории вероятностей и математической статистики.

Для достижения цели исследования поставлены следующие задачи:

1. Провести анализ требований к данным для существующих подходов к косвенному оцениванию искомой характеристики – кумулятивной интенсивности поведения индивидов.

2. Определить при помощи вычислительного эксперимента точность существующих подходов к косвенному оцениванию интенсивности поведения. Так как существующие подходы к оцениванию характеризуют искомое распределение с разных точек зрения: одни позволяют получить оценку параметра, а другие – подгонку самого распределения, то в качестве показателя точности была использовано расстояние Канторовича-Рубинштейна (метрика Вассерштейна) между вероятностными распределениями. Для достижения цели исследования предлагается рассмотреть несколько значений параметров исходного смешивающего распределения, которые отвечают обыденному поведению индивидов.

3. На основе сравнения существующих подходов по результатам пунктов 1 и 2 сформулировать рекомендации для применения конкретных методов в рамках в зависимости от исходных данных.

Статья организована следующим образом. Во втором разделе представлено описание существующих вероятностных подходов оценивания интенсивности поведения. В основе рассматриваемых подходов лежит смешанный пуассоновский процесс и его свойства, которые описаны в разделе 2.1. Этот раздел служит также для введения терминов, которые используются в дальнейшем при описании подходов и сравнении их между собой. Существующие подходы косвенной оценки характеристик поведения можно разделить на две группы: опирающиеся на оценки параметра смешивающего гамма-распределения (раздел 2.2) и опирающиеся на методы приближения смешивающего распределения с использованием байесовских сетей доверия (раздел 2.3). В соответствующих подразделах приведены схемы получения оценки, особое внимание уделялось объему наблюдений, которые требуются для построения оценки, так как каждый из подходов может использовать различные методы, и, соответственно, допускает вариабельность в исходных предположениях. Раздел 3 содержит описание вычислительного эксперимента, которые служит для определения точности рассматриваемых методов. Для проведения

эксперимента был сгенерирован набор данных, который отражает структуру гамма-пуассоновской модели поведения для нескольких наборов параметров, отвечающих обыденному поведению индивидов. С использованием этого набора данных были определены косвенные оценки интенсивности поведения с использованием рассматриваемых подходов, затем эти оценки сравнивались с истинным значением параметра. Так как два блока подходов различаются по сути получаемой косвенной оценки (оценка параметра распределения и подгонка распределения), то в качестве меры сходства было выбрано расстояние Канторовича-Рубинштейна. Раздел 4 содержит обсуждение проведенного эксперимента и применимости рассматриваемых в исследовании методов. Практической значимостью обладают сформулированные в разделе рекомендации по применению методов в зависимости от характеристик ситуации, в которой требуется получить косвенную оценку интенсивности. Отметим, что ранее такая работа не проводилась. Научный вклад исследования состоит в систематизации существующих подходов к косвенному оцениванию интенсивности поведения, а также в представлении качественных и количественных характеристик их применимости в ситуации ограниченных наблюдений.

2. Существующие подходы к оцениванию интенсивности поведения на основе ограниченных данных. Как уже отмечалось, необходимость использование математических моделей для построения оценок характеристик поведения возникает в силу неопределенности доступных данных и знаний о поведении. Так, в рамках одномоментного среза (опроса), индивид может предоставить информацию лишь о самых недавних или же о самых запоминающихся эпизодах [3], при этом оценка совокупных (кумулятивных) характеристик таких последовательностей эпизодов в группах индивидов используется при мониторинге социально значимого поведения. Автоматизированные системы обработки такой информации о поведении опираются на математические модели поведения, которые позволяют моделировать возникающую неопределенность. С вероятностной точки зрения базовым математическим объектом для таких моделей является пуассоновский процесс и его вариации [11].

2.1. Гамма-пуассоновская модель поведения. Естественной математической моделью для последовательностей риск-ассоциированных эпизодов являются точечные случайные процессы и гамма-пуассоновский процесс в частности [10, 11]. Смешивающее распределение моделирует индивидуальные особенности при оценке кумулятивных характеристик поведения в группах индивидов.

Такая неоднородность (гетерогенность) возникает в силу некоторых ненаблюдаемых индивидуальных особенностей, которые носят название *склонности к событиям* или *склонности к риску* [10].

Пусть рассматриваются последовательности эпизодов некоторого интересующего нас поведения в группе из n индивидов. Предполагается, что рассмативаемые последовательности удовлетворяют предположениям гамма-пуассоновской модели поведения [12]: эпизоды для каждого индивида в отдельности могут быть описаны процессом Пуассона, а индивидуальные особенности – некоторой случайной величиной u_i, i=1. . . n, для наблюдаемых индивидов u_i – независимые и одинаково распределенные случайные величины. В этом случае последовательность эпизодов, реализующаяся в группе индивидов полностью характеризуется функцией интенсивности:

$$\lambda(t|u_i) = u_i \rho_i(t), i = 1 \dots n,$$

здесь u_i моделируют индивидуальную склонность к реализации эпизодов для наблюдаемых объектов, а $\rho_i(t)$ – возможное влияние времени и внешних ковариант на реализацию эпизодов процесса. Если нет иных предположений о природе таких индивидуальных особенностей, то удобно полагать, что u_i имеют гамма-распределение вероятности с одним параметром $\phi > 0$:

$$f(s;\phi) = \frac{s^{\phi^{-1}-1}\exp(-s/\phi)}{\phi^{\phi^{-1}}\Gamma(1/\phi)}, s > 0. \quad (1)$$

Таким образом, исследуемый паттерн реализации эпизодов в группе неоднородных индивидов, описывается гамма-распределением величины, моделирующей эту неоднородность, или просто параметрами этого распределения ϕ. Существуют несколько подходов к оценке искомого гамма-распределения вероятности, которую обозначим g_u^{est}:

– подход регрессии Кокса;

– подход на основе классических БСД, который требует дискретизации входных данных модели;

– подход на основе гибридных БСД, который позволяет отказаться от дискретизации;

– оценка параметра копулы (который является искомым параметром функции интенсивности).

Для построения оценок используется доступная информация об эпизодах рассматриваемого поведения: для каждого респондента

i известна информация о датах m последних последовательных эпизодов, т.е. определены длины $m-1$ интервалов между последовательными эпизодами поведения $W_j^i, j = 1 \dots (m-1)$.

2.2. Методы оценки параметра смешивающего гамма-распределения в гамма-пуассоновской модели поведения. **Регрессионная модель Кокса** является классическим методом, использующимся при анализе процессов повторяющихся событий [11]. В этом случае в качестве зависимой переменной регрессии выступает функция риска, которая тесно связана с функцией интенсиности процесса [11]:

$$h(B(t)) = u_i \rho_i(t) = u_i \rho_0(t;\alpha)\exp(x_i(t)\beta),$$

где $B(t)$ – интервал времени с момента последнего эпизода, $\rho_0(t;\alpha)$ – базовая функция, отражающая поведение функции интенсивности при отсутствии воздействия ковариант, $x_i(t)$ – коварианты процесса, β –вектор коэффициентов регрессии.

Подгонка регрессии осуществляется при помощи метода максимального правдоподобия, и в случае включения в модель случайного эффекта, обуславливающего индивидуальные различия в форме гамма-распределения вероятности (1), строится также оценка его параметра ϕ. Однако, как показывают вычислительные эксперименты [13], мощность статистического теста на равенство параметра гамма-распределения нулю невысока, особенно если для каждого индивида наблюдаются менее 20 эпизодов. Отметим также, что в этом случае оценка максимума правдоподобия оказывается смещенной.

Применимость подхода обусловлена, с одной стороны, распространенностью соответствующего программного обеспечения, и, с другой стороны, тем, что этот метод опирается на подходы анализа времени жизни, которые широко применяются при оценке риска и язык которых известен в этой сфере деятельности. Также этот подход позволяет учитывать внешние факторы и их зависимость только в форме линейной регрессии.

Однако согласно [14], чтобы параметр регрессии был идентифицируемым, необходимо наблюдение нескольких ковариант процесса эпизодов, и потому в вычислительном эксперименте этот подход не участвует. Подход регрессии Кокса может использоваться в ситуациях, когда отсутствует нехватка данных и кроме самих эпизодов поведения наблюдаются иные факторы, которые объясняют вариабельность между индивидами.

Оценка параметра копулы, связывающей длины интервалов между последовательными эпизодами поведения в гамма-пуассоновской модели, является параметрическим методом получения искомого параметра. В основе этого подхода лежит свойство смешанного гамма-пуассоновского процесса, которое утверждает, что копулой зависимости нескольких длин интервалов между последовательными интервалами [15] выступает *копула Клейтона* [16].

В этом случае функцию их совместного распределения можно представить в виде [17]:

$$P(W_1^i > w_1, \ldots, W_m^i > w_m) = \left(\sum_{j=1}^{m} S_j^i(w_j)^{-\phi} - (m-1) \right)^{-\phi^{-1}} = \\ = C_{\text{clayton}}(S_1^i(w_1), \ldots, S_m^i(w_m)),$$

где $S_j^i(w_j) = P(W_j^i > w_j)$ – маржинальные распределения, ϕ – параметр гамма-распределения (1), $C_{\text{clayton}}(\cdot)$ – функция, объединяющая маржинальные распределения в совместное распределение вероятности

$$C_{\text{clayton}}(z_1, \ldots, z_m) = \Psi(\Psi(z_1) + \ldots + \Psi(z_m)),$$

$$\Psi(z) = \frac{1}{\phi}(z^{-\phi} - 1).$$

Таким образом, оценка параметра копулы, связывающей длины наблюдаемых интервалов между эпизодами поведения, является искомой оценкой параметра гамма-распределения вероятности u_i. Существуют различные методы построения оценки параметра копулы, в том числе классический метод максимальное правдоподобия или обращения рангового коэффициента корреляции, и для небольших выборок возможно смещение оценки. Отмечается [18], классические методы позволяют получить достоверные оценки, если наблюдается от 25 индивидов; однако для оценки параметра копулы Клейтона были предложены методы [19], позволяющие оценивать параметр всего по 10 наблюдениям. Особенностью применения этого метода является необходимость перехода к псевдо-наблюдениям, получаемым посредством интегрального преобразования вероятности.

Последовательность шагов построения оценки параметра при помощи аппарата копул можно сформулировать при помощи следующей последовательности шагов.

1. Собрать данные об m эпизодах для n объектов наблюдения, причем число m должно быть одинаковым для всех, $m \geqslant 2$ (в вычислительном эксперименте используется $m = 2$). Число объектов наблюдения может быть относительно невелико, порядка 10 или 50, и зависит от конкретного метода оценивания параметра копулы на последнем шаге. Собранные данные представляют собой набор длин интервалов между последовательными эпизодами поведения W_j^i.

2. Привести значения сформированной на шаге 1 выборки к равномерному распределению при помощи преобразования:

$$u_j^i = r_j^i/(n+1), i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\},$$

где r_j^i обозначает ранг наблюдения W_j^i среди всех $i \in 1, ..., n, j \in 1, ..., m$.

3. Оценить значение параметра функции Клейтона при помощи полупараметрического метода обращения коэффициента корреляции Кендалла [18] (или любым иным методом).

4. Для сравнения на основе полученной оценки параметра были сгенерированы 1000 реализаций гамма-распределенной случайной величины.

Полученная оценка является оценкой искомого параметра смешивающего гамма-распределения вероятности. В вычислительном эксперименте на основе этой оценки были сгенерированы значения гамма-распределенной случайной величины.

2.3. Методы приближения смешивающего распределения с использованием байесовских сетей доверия. Однако методы, опирающиеся на оценку параметра, обладают ограниченными возможностями для обработки данных с неопределенностью, которые часто возникают при обращении к самоотчетам и интервью [20]. Гибким инструментом для работы с такими данными являются БСД.

Классическая БСД в качестве модели оценивания интенсивности по ограниченным данным была предложена в работах [21,22]. Графическая структура сети формируется следующим образом: центральным узлом является переменная, моделирующая различия между индивидами в выборке u_i (она же позволяет вычислить среднее число эпизодов в промежутке времени), а детьми этого узла являются узлы, сопоставленные переменным-длинам интервалов между последовательными эпизодами. Таким образом, возникает звездчатая структура, которая обуславливает

разбиение совместного распределения вероятности (например, для трех интервалов между последовательными эпизодами поведения):

$$P(u^i < u, W_1^i < t_1, W_2^i < t_2, W_3^i < t_3) = P(u^i < u)P(W_1^i < t_1 \mid u^i) \\ P(W_2^i < t_2^i \mid u^i)P(W_3^i < t_3^i \mid u^i).$$

Подгонка искомого гамма-распределения в рамках этого подхода осуществляется с помощью проведения апостериорного вывода с использованием ранее опредленных параметров сети при заданной экспертным образом структуре сети [21]. Оценка параметров БСД осуществляется при помощи метода максимального правдоподобия [23, 24]. При этом априорные параметры модели могут быть заданы на основе статистической информации или же предоставлены экспертами.

Этот подход широко применяется в различных ситуациях, связанных с неопределенностью данных [24]. Подход классических БСД обладает высокой применимостью в приложениях благодаря развитому программно-аналитическому обеспечению процесса байесовского рассуждения, структура БСД отражает интуитивно понятные причинно-следственные связи в предметной области. Основным его недостатоком для построения оценок в рамках гамма-пуассоновской модели поведения является необходимость дискретизации поступающих данных о длинах инетрвалов между последовательными эпизодами, что значительно увеличивает число параметров, которое требуется оценить/задать для модели. Например, если для оценки функции интенсивности гамма-пуассоновского процесса используются k переменных, отвечающих длинам интервалов между последовательными эпизодами, каждая из которых разбита на 5 интервалов, то общее число параметров БСД для ядра гамма-пуассоновской модели, которые необходимо или оценить или задать экспертным путем составляет $24 * k$, что представляет большую нагрузку на данные или экспертов.

В качестве оценки искомого смешивающего распределения выступает апостериорное дискретное распределение вероятности, получаемое при пропагации свидетельств о длинах интервалов между эпизодами, и объем выборки не важен вследствие использования байесовского подхода.

Подход гибридных БСД обладает гибкостью классических БСД к моделированию неопределенности, и при этом отражает особенности гамма-пуассоновской модели эпизодического процесса,

как непрерывность составляющих ее переменных. Возможны различные подходы к заданию гибридной БСД, далее рассмотрено приближение дискретных распределений при помощи смесей усеченных базисных функций [25, 26]. Основной идеей метода является приближение совместной плотности n-мерного вектора $\mathbf{X}$, отражающего распределение переменных в БСД, при помощи наборов вещественных функций $\Psi = \{\psi_i\}_{i=0}^{\infty}, \psi_i : \mathbb{R} \to \mathbb{R}$:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{k} \prod_{j=1}^{c} a_{i,\mathbf{y}}^{(j)} \psi_i(z_j),$$

где $\mathbf{Y} = (Y_1, \ldots, Y_d)$ дискретная часть вектора $\mathbf{X}$, $\mathbf{Z} = (Z_1, \ldots, Z_c)$ – непрерывная часть; $a_{i,\mathbf{y}}^{(j)}$ вещественные числа. В качестве базисных функций ψ_i часто используются полиномы $\psi_i(x) = x^i$ и экспоненты $\Psi(x) = \{1, \exp(-x), \exp(-2x) \ldots\}$.

Для рассмотренной выше структуры БСД для гамма-пуассоновской модели поведения с тремя интервалами между последовательными эпизодами поведения, разложение с помощью усеченных экспонент выглядит следующим образом:

$$\hat{f}(u, t_1, t_2, t_3) = a_0 + \sum_{i=1}^{m} a_i \exp\left(b_i^{(u)} u + b_i^{(W_1)} t_1 + b_i^{(W_2)} t_2 + b_i^{(W_3)} t_3\right), \quad (2)$$

где m представляет собой глубину приближения базисными функциями.

При задании априорных параметров модели этот подход является достаточно требовательным к данным, однако как и все модели на основе БСД, модель может быть задана один раз, и затем использована для построения оценок интенсивности поведения. При использовании этого подхода оценка интенсивности поведения представляет собой приближенное распределение вероятности, получаемое с помощью пропагации свидетельств о длинах интервалов между последовательными эпизодами поведения.

Сама процедура численного задания исходной модели опирается на решение задачи квадратичной оптимизации для определения параметров смесей, описывающих маргинальные и условные распределения, отвечающие графу БСД [26]. Количество параметров таких смесей может варьироваться, соответственно, варьируется и объем необходимых

данных. Кроме того, алгоритмы приближения могут опираться на априорные знания экспертов.

Последовательность шагов построения косвенной оценки при помощи байесовских сетей доверия можно сформулировать при помощи следующей последовательности шагов.

1. Задать структуру зависимости наблюдаемых переменных, ядром которой является звездчатая структура гамма-пуассоновской модели. На этом этапе возможно включение в модель факторов внешней среды, недоопределенность знаний о их влиянии может быть квантифицирована экспертно или же статистически.

2. Сформировать тренировочный набор данных, который содержит достаточно большое число наблюдений интервалов между последовательными эпизодами исследуемого поведения и оцениваемой частоты, а также иные наблюдаемые факторы. В рамках вычислительного эксперимента сгенерированный набор данных был разбит на тестовую и тренировочную выборки в отношении 2:1.

3. (шаг для применения подхода классической БСД) Дискретизовать непрерывные данные об интервалах между последовательными эпизодами W_j^i на большое число интервалов. Для проведения вычислительного эксперимента использовалось разбиение на 8 интервалов при помощи метода взаимной информации Хартеминка [27].

4. При помощи специализированных алгоритмов (для смесей усеченных базисных функций [26], для классических байесовских сетей [27]) обучить параметры байесовской сети доверия на тестовой выборке. В исследовании использовался метод максимума правдоподобия.

5. Провести байесовский вывод для тестового набора поступающих данных (произвольного объема, от одного наблюдения) для получения апостреиорного распределения искомой переменной ϕ, отражающей смешивающее распределение гамма-пуассоновского процесса. В рамках вычислительного эксперимента для получения апостериорного распределения использовалась тестовая выборка сгенерированного тестового набора данных.

Существуют и иные подходы к работе с непрерывными распределениями в рамках формализма БСД, позволяющие использовать меньшее количество наблюдений для оценки априорных параметров, например, подход лозы [28]. В таблице 2 приведен сравнительный анализ существующих подходов с точки зрения применимости в условиях ограниченных ресурсов. Проанализированы возможности моделей к учету внешних факторов, влияющих на поведение, требования к данным

для задания той или иной модели, а также число параметров модели оценивания.

3. Точность методов оценивания интенсивности поведения в гамма-пуассоновской модели поведения по ограниченным данным. Анализ поведения человека является трудоемкой задачей, сбор данных в которой часто связан со значительными сложностями. Так, для прямой оценки интенсивности рискованного поведения используют ресурсозатратные способы прямого наблюдения или же дневниковые методы, которые не могут охватить все интересующие виды поведения [3]. Поэтому использование сгенерированных данных часто является необходимым шагом при оценке численных показателей автоматизированных систем обработки информации о поведении.

Особого внимания требует вопрос о релевантности ключевого объекта работы: гамма-пуассоновской модели поведения, реально наблюдаемому поведению индивидов. Эта модель опирается на классические предпосылки пуассоновского процесса (подробнее этот вопрос освещен в работе [12]), однако модели оценивания кумулятивной характеристики, рассматриваемые в работе, могут быть адаптированы к возможным отклонениям от этих предполсылок. Так, функция интенсивности процесса может включать зависимость от различных ковариант в форме регрессии [11], и подобные модели широко используются при анализе риска. Аналогично, использование БСД допускает внедрение узлов, которые могут аккумулировать отклонения от исходных предположений. Отметим также, что в ситуациях острой ограниченности ресурсов, когда доступны данные лишь о нескольких эпизодах для небольшого числа индивидов, нет возможности установить принадлежность наблюдаемых эпизодов тому или иному классу случайных процессов. Рассматривые в работе методы позволяют получить предварительную косвенную оценку уже на сверхмалых выборках, что важно на ранних этапах мониторинга поведения индивидов.

Для достижения цели исследования был проведен вычислительный эксперимент, в основе которого лежит генерация набора данных, отвечающего свойствам гамма-пуассоновской модели поведения, согласно шагам псевдокода 1. Его шаги отражают структуру гамма-пуассоновской модели поведения (раздел 2.1): сначала генерируется значение ненаблюдаемой случайной интенсивности процесса (шаг 11), имеющее гамма-распределение вероятности; затем на основе этого значения генерируются пять случайных величин, имеющие экспоненциальное распределение вероятности (шаги 12-16), на основании которых рассчитывались значения оценок с помощью различных подходов. Для

применения подхода на основе копул сгенерированные значения были отнормированы согласно рангам (шаг 21, шаги 1-6).

Для достижения цели работы использовались несколько значений исходного гамма-распределения, которые наиболее характерны для повседневного эпизодического поведения (шаг 7). Объем датасета составил 10000 наблюдений, что удовлетворяет требованиям существующих подходов к оценке, которые описаны в разделе 2. При генерации случайных величин были отброшены очень длинные интервалы, которые естественно возникают при обращении к пуассоновскому процессу с низкой интенсивностью, однако наблюдение которых в рамках самоотчета крайне маловероятно (функция 1-7 и шаги 22-27).

Листинг 1. Последовательность шагов генерации набора данных, отражающих стурктуру гамма–пуассоновской модели

```
1: function Truncate(Data)
2:     NewData ← Data
3:     for i ∈ (1 : NumberOfColumns(Data) − 1) do
4:         NewData[, i + 1] ← Data[which(Data[, i + 1] < 1000), i + 1]
5:     end for
6:     return NewData
7: end function
8: S ← [0.2, 0.5, 0.8, 1.0, 1.2, 1.5]
9: SimData ← DataFrame[nrow = 10000, ncol = 36]
10: for s ∈ S do
11:     for i ∈ [1, 10000] do
12:         λ_s[i] ← RandomGamma(s, 1/s)
13:         τ_s^1[i] ← RandomExponential(λ)
14:         τ_s^2[i] ← RandomExponential(λ)
15:         τ_s^3[i] ← RandomExponential(λ)
16:         τ_s^4[i] ← RandomExponential(λ)
17:         τ_s^5[i] ← RandomExponential(λ)
18:     end for
19:     MiniData ← ColumnBind(λ_s, τ_s^1, τ_s^2, τ_s^3, τ_s^4, τ_s^5)
20:     SimData ← ColumnBind(SimData, MiniData)
21: end for
22: data1 ← Truncate(simdata[, 1 : 6])
23: data2 ← Truncate(simdata[, 7 : 12])
24: data3 ← Truncate(simdata[, 13 : 18])
25: data4 ← Truncate(simdata[, 19 : 24])
26: data5 ← Truncate(simdata[, 25 : 30])
27: data6 ← Truncate(simdata[, 31 : 36])
```

С использованием сгенерированного набора данных, в среде статистической обработки данных R были получены оценки искомого гамма-распределения для разного количества наблюдений (20, 40, 60, 80) трех существующих подходов:

1. $g_u^{\text{ClassicBBN}}$ для подхода, подразумевающего получение апостериорного распределения на основе классических БСД; для

получения оценок использовался алгоритм взвешенного правдоподобия, реализованный в пакете bnlearn [27];

2. g_u^{copula} для подхода, в основе которого лежит оценка параметра копулы, связывающей две переменные – длины интервалов в гамма-пуассоновской модели поведения; использовался классический алгоритм обращения рангового коэффициента корреляции, реализованный в пакете copula [1];

3. $g_u^{\text{HybridBBN}}$ для подхода, подразумевающего получение эмпирического апостериорного распределения на основе гибридных БСД, где непрерывные распределения приближены с помощью смесей усеченных экспонент, реализованные в пакете MoTBFs [2].

При этом все рассматриваемые методы являются различными по сути получаемых оценок. Метод на основе копул выдает оценку параметра смешивающего гамма-распределения, и потому итогом является непрерывное распределение вероятности. Ранее разработанный подход с ядром в виде классической БСД дает оценку в виде дискретного распределения вероятности с небольшим числом ступеней, в то время как гибридные БСД, приближенные с помощью смесей усеченных распределений, позволяют получить эмпирическое непрерывное распределение неизвестного класса. В силу разнообразной природы оценок, получаемых в различных моделях, для сравнения их точности было решено использовать расстояние Канторовича-Рубинштейна (метрику Вассерштайна), которая является мерой удаленности двух вероятностных распределений друг от друга. Метрика Канторовича-Рубинштейна [29] представляет собой естественное расстояние на пространстве распределений вероятности.

В диссертационном исследовании вычисляется следующим образом. Пусть a и b представляют собой значения некоторой случайной величины, тогда их эмпирические функции распределения есть $F(t) = \sum_{i=1}^{m} w_i^{(a)} \mathbb{1}\{a_i \leqslant t\}$ и $G(t) = \sum_{j=1}^{n} w_j^{(b)} \mathbb{1}\{b_j \leqslant t\}$ ($w_i^{(a)}$ и $w_j^{(b)}$ представляют собой веса). Расстояние между такими функциями может быть вычислено как

$$W_p(F, G) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p}.$$

[1] Hofert M., Kojadinovic I., Maechler M. and Yan J. (2023). copula: Multivariate Dependence with Copulas. R package version 1.1-3 URL:https://CRAN.R- project.org/package=copula

[2] Pérez-Bernabé I., Salmerón A., Nielsen T.D., Maldonado A.D. (2022). MoTBFs: Learning Hybrid Bayesian Networks using Mixtures of Truncated Basis Functions. R package version 1.4.1, https://CRAN.R- project.org/package=MoTBFs

Здесь F^{-1} и G^{-1} – обобщенные обратные функции. Для $p = 1$ также выполнено:

$$W_1(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|dx.$$

Таблица 1 отражает рассчитанные расстояния Канторовича-Рубинштейна для различных моделей оценивания и различных значений исходных параметров. Расстояние вычислялось с использованием пакета transport[3].

4. Обсуждение. Основой целью работы является определение точности существующих методов оценивания интенсивности поведения. В таблице 1 представлены результаты вычислительного эксперимента, проведенного для оценки точности существующих подходов к оцениванию искомой характеристики. Отметим, что самые точные косвенные оценки позволяет получить подход, опирающийся на оценку параметра копулы, который при этом не позволяет напрямую учитывать контекст реализации эпизодов. Более гибкий подход БСД позволяет получать достаточно точные оценки при использовании неперерывных переменных в узлах; классический же подход дает более далекие от истинных значения.

Точность получаемой оценки является одним из ключевых факторов, обуславливающих применимость существующих моделей оценивания кумулятивной характеристики эпизодического поведения в группе индивидов, различающихся по склонности к поведению. Кроме этого важно учитывать нагрузка на данные, т.е. сколько данных необходимо использовать для получения оценки. С другой стороны важно учитывать контекст реализации эпизодов и возможное влияние внешних факторов.

Таблица 2 обобщает результаты работы, и представляет возможность выбора подхода в зависимости от имеющихся в распоряжении исследователя данных и целей его исследования. Качественное сравнение подходов к построению оценки опирается на описание вводных данных, представленное в 2 (пункт 1 для алгоритма оценки параметра копулы и пункты 1–2 для алгоритма построения приближенного распределения на основе БСД). Под относительной точностью понимается отношение значения расстояния Канторовича-Рубинштейна между оценкой, получаемой тем или иными методом, и истинным значением и расстояния между оценкой, получаемой методом

[3] Schuhmacher D., Bähre B., Bonneel N., Gottschlich C., Hartmann V., Heinemann F., Schmitzer B., Schrieber J. (2024). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.15-2. https://cran.r- project.org/package=transport

классической БСД, и истинным значением в случае, когда оценка строится по 20 наблюдениям для различных значений параметров гамма распределения. m – число рассматриваемых эпизодов, k – число интервалов дискретизации.

Таблица 1. Среднее значение и 95% бутстреп доверительный интервал для расстояния Канторовича-Рубинштейна между оценкой частоты, полученной при помощи различных подходов по n наблюдениям, и истинным смешивающим гамма-распределением, для 500 репликаций исходной выборки

Модель оценивания	Значение параметра исходной модели					
	$\phi = 1.5$	$\phi = 1.2$	$\phi = 1$	$\phi = 0.8$	$\phi = 0.5$	$\phi = 0.2$
Классическая БСД (исходная модель)						
$n = 20$	0.55, (0.18, 1.3)	0.94, (0.28, 1.8)	0.87, (0.24, 1.85)	1.12, (0.3, 2.48)	1.37, (0.33, 2.85)	2.96, (0.56, 6.35)
$n = 40$	0.49, (0.2, 0.99)	0.88, (0.36, 1.62)	0.81, (0.29, 1.44)	1.17, (0.47, 2.03)	1.30, (0.42, 2.54)	2.83, (0.83, 5.17)
$n = 60$	0.48, (0.22, 0.87)	0.87, (0.41, 1.43)	0.78, (0.33, 1.34)	1.12, (0.5, 1.92)	1.26, (0.48, 2.24)	2.84, (1.21, 4.85)
$n = 80$	0.46, (0.23, 0.75)	0.88, (0.45, 1.36)	0.78, (0.35, 1.27)	1.12, (0.6, 1.76)	1.28, (0.56, 2.13)	2.88, (1.51, 4.42)
Оценка параметра копулы						
$n = 20$	0.21, (0.03, 0.52)	0.20, (0.03, 0.52)	0.20, (0.03, 0.57)	0.20, (0.04, 0.51)	0.19, (0.04, 0.48)	0.30, (0.07, 0.62)
$n = 40$	0.14, (0.03, 0.38)	0.14, (0.03, 0.36)	0.14, (0.03, 0.36)	0.14, (0.03, 0.37)	0.14, (0.04, 0.35)	0.29, (0.09, 0.52)
$n = 60$	0.12, (0.03, 0.31)	0.12, (0.03, 0.32)	0.12, (0.03, 0.31)	0.11, (0.03, 0.26)	0.13, (0.04, 0.29)	0.28, (0.11, 0.49)
$n = 80$	0.11, (0.03, 0.27)	0.10, (0.03, 0.25)	0.10, (0.03, 0.26)	0.10, (0.03, 0.25)	0.12, (0.04, 0.29)	0.29, (0.13, 0.46)
Гибридная БСД (приближение усеченными базисными функциями)						
$n = 20$	0.26, (0.12, 0.54)	0.24, (0.12, 0.51)	0.26, (0.13, 0.53)	0.33, (0.15, 0.73)	0.45, (0.19, 1.03)	0.94, (0.35, 2.66)
$n = 40$	0.19, (0.09, 0.39)	0.18, (0.08, 0.38)	0.19, (0.1, 0.37)	0.25, (0.11, 0.52)	0.35, (0.14, 0.86)	0.84, (0.28, 2.12)
$n = 60$	0.16, (0.08, 0.32)	0.15, (0.07, 0.29)	0.16, (0.08, 0.3)	0.20, (0.09, 0.42)	0.30, (0.13, 0.65)	0.76, (0.26, 1.68)
$n = 80$	0.14, (0.07, 0.26)	0.13, (0.06, 0.24)	0.14, (0.08, 0.25)	0.18, (0.08, 0.37)	0.27, (0.12, 0.62)	0.78, (0.24, 1.63)

Анализ таблицы 2 позволяет сказать, что если не предполагается учитывать внешние факторы, то меньше всего требований к данным предъявляется при использовании подхода к оцениванию параметра искомого гамма-распределения при помощи копул. В этом случае существуют методы, которые позволяют строить оценку параметра копулы Клейтона всего по 10 наблюдениям.

Если говорить о ситуациях, когда важно учитывать влияние внешних факторов, то в этом случае используют подходы на основе БСД, которые после полного задания априорных значений параметров позволяют получать подгонку искомого гамма-распределения вероятности по любому количеству имеющихся данных и наблюдаемых интервалов для индивидов. Однако для использования байесовского рассуждения необходимо полностью задать модель БСД. Структура сети определена ранее, и различные подходы к заданию параметров БСД имеют различное число параметров к оценке. В целом, обе модели имеют сопоставимое число априорных параметров для задания. В случае классической БСД эти параметры представляют собой значения в таблицах условных

вероятностей, соответствующие ребрам, которые соединяют ключевую переменную и переменные-длины интервалов между эпизодами. Таких переменных будет $(m^2 - 1)$ для каждого узла. В случае гибридной БСД оцениваются параметры линейной комбинации базисных функций из разложения в формуле (2), и финальное число параметров зависит от числа элементов этого разложения. В целом, подгонка такого разложения требует большого числа наблюдений, порядка 30 для каждого из m параметров в (2).

Таблица 2. Сравнительный анализ методов оценивания интенсивности поведения с точки зрения результативности их применения в условиях ограниченных данных

Модель	Возможности учета влияния внешних факторов	Возможности обработки неточности поступающих значений	Требования к данным	Относительная точность получаемой оценки*
Параметрические подходы				
Регрессия пропорциональных рисков (Кокса)	Только в форме линейной регрессии	Нет	Более 20 наблюдаемых эпизодов для каждого индивида + наличие наблюдаемых факторов	–
Оценка параметра копулы	Нет	Нет	Оценивается 1 параметр; есть алгоритмы оценки для 10 и более наблюдений	0.21
Байесовские подходы				
БСД: классическая дискретизация	Да	Да	Задание модели происходит однократно на большом наборе данных или экспертно $(m*(k^2-1)$ параметров); затем для получения оценки требования к данным отсутствуют	1
Гибридная БСД: приближение смесями базисных функций	Да	Да	Задание модели происходит однократно на большом наборе данных; затем для получения оценки требования к данным отсутствуют	0.33

Отметим также, что в целом различные особенности входных данных могут учитываться моделями на основе БСД. Однако, поступающие данные, которые содержат различное количество эпизодов для разных индивидов, может учитывать также и подход регрессии Кокса.

Таким образом можно заключить, что переход к неперерывным моделям оценивания гамма-распределения позволяет получать распределение вероятности, более близкое к истинному значению, чем распределение, получаемое с помощью классической БСД. Можно

сформулировать следующие рекомендации по применению имеющихся моделей построения косвенной оценки интенсивности поведения:

– Если наблюдений много, и для каждого индивида наблюдается много эпизодов поведения, то подход *регрессии Кокса* является предпочтительным в силу развитого аппарата.

– Если наблюдений много, но о каждом индивиде известны только несколько эпизодов поведения, которые могут быть сопряжены с неопределенностью (например, получены в результате самоотчета), то подход *гибридных БСД* является предпочтительным в силу гибкости к учету возникающей неопределенности данных и знаний. Кроме того, впоследствии построенная модель гибридной БСД может использоваться для поддержки принятия решений в условиях острой нехватки наблюдений.

– Если же наблюдается небольшое число индивидов и несколько эпизодов для каждого из них, то достаточно точную численную оценку можно получить на основе оценки параметра копулы. Однако если при этом необходимо учитывать дополнительные факторы, то возможно использование гибридных БСД, заданных с помощью экспертов [30].

Среди ограничений проведенного исследования можно отметить небольшой разброс параметров гамма-распределения, которые использовались для формирования набора данных. Это обусловлено тем, что рассматривали те параметры, которые могут встречаться при моделировании обыденного поведения. Также в исследовании не были рассмотрены иные модели задания гибридных БСД.

5. Заключение. Для получения кумулятивных характеристик последовательностей эпизодов в группах индивидов существуют несколько подходов, которые опираются на математическую модель процесса событий: два из них опираются на оценку параметра искомого распределения, два – на предварительно заданную БСД. Для достижения цели исследования были решены:

1. Проведен анализ требований к данным для существующих подходов к косвенному оцениванию искомой характеристики. Основные результаты представлены в таблице 2. В результате анализа было установлено, что большими возможностями к учету возможного влияния сопутствующих факторов обладают подходы, основанные на БСД. Наименьшее число параметров требуется оценить при использовании параметрического подхода на основе оценки параметра копулы.

2. При помощи вычислительного эксперимента определена точность существующих подходов к косвенному оцениванию интенсивности поведения. Значения расстояния Канторовича-

Рубинштейна, полученные в рамках эксперимента, представлены в таблице 1. Показано, что для типовых значений параметров подходы, опирающиеся на непрерывные данные, обладают более низкими значениями расстояния Канторовича-Рубинштейна.

3. В разделе 4 были сформулированы рекомендации для применения конкретных методов в рамках в зависимости от исходных данных.

В работе проанализирована близость получаемой оценки к истинному значению в терминах расстояния Канторовича-Рубинштейна. Наивысшей точностью в сочетании с наиболее низкими требованиями к количеству наблюдаемых индивидов обладает подход на основе оценки параметра копулы, связывающей длины интервалов между последовательными эпизодами в гамма-пуассоновской модели поведения. При этом подход на основе копул не приспособлен для учета различной неопределенности, часто сопутствующей ситуации сбора самоотчетов.

Практической значимостью обладают сформулированные в таблице 2 особенности применимости различных подходов к оцениванию искомой характеристики. Этот результат является новым в сфере создания модельно-алгоритмического обеспечения систем обработки информации о поведении, и позволяет осуществлять выбор подходящего метода оценивания исходя из имеющихся данных. Дальнейшим направлением исследований является расширение вычислительного эксперимента для определения точности оценки гамма-распределений с иными параметрами, а также исследование точности оценок в ситуации ограниченных данных.

Литература

1. Chavez K., Palfai T.P., Cheng D.M., Blokhina E., Gnatienko N., Quinn E.K., Krupitsky E., Samet J.H. Hazardous Alcohol Use, Impulsivity, and HIV-Risk Behavior Among HIV-Positive Russian Patients With a History of Injection Drug Use // The American journal on addictions. 2021. vol. 30. no. 2. pp. 164–172.
2. Hendrieckx C., Ivory N., Singh H., Frier B.M., Speight J. Impact of severe hypoglycaemia on psychological outcomes in adults with type 2 diabetes: a systematic review // Diabetic Medicine. 2019. vol. 6. no. 9. pp. 1082–1091.
3. Пащенко А.Е., Тулупьев А.Л., Тулупьева Т.В., Красносельских Т.В., Соколовский Е.В. Косвенная оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения // Здравоохранение Российской Федерации. 2010. № 2. С. 32–35.
4. Wojciechowski T.W. Major depressive disorder as a moderator of the relationship between heavy-episodic drinking and anxiety symptoms // Journal of mental health. 2023. pp. 1–8. DOI: 10.1080/09638237.2023.2245889.
5. Lewer D., Freer J., King E., Larney S., Degenhardt L., Tweed E.J., Hope V., Harris M., Millar T., Hayward A., Ciccarone D., Morley K. Frequency of health-care utilization by

adults who use illicit drugs: a systematic review and meta-analysis // Addiction. 2020. vol. 115. no. 6. pp. 1011–1023.

6. Feldhege J., Moessner M., Bauer S. Who says what? Content and participation characteristics in an online depression community // Journal of Affective Disorders. 2020. vol. 263. pp. 521–527.
7. Jiotsa B., Naccache B., Duval M., Rocher B., Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness // International journal of environmental research and public health. 2021. vol. 18. no. 6. DOI: 10.3390/ijerph18062880.
8. Олисеенко В.Д., Хлобыстова А.О., Корепанова А.А., Тулупьева Т.В. Автоматизация оценки темперамента пользователей онлайн социальной сети // Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514. № 2. С. 235–241. DOI: 10.31857/S2686954323601471.
9. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users // Intelligent Distributed Computing XIII. Springer International Publishing, 2020. pp. 272–277.
10. Grandell J. Mixed Poisson Processes. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC. 1997. 280 p.
11. Cook R.J., Lawless J.F. The statistical analysis of recurrent events. Springer New York, 2007. 404 p.
12. Stoliarova V.F., Tulupyev A.L., Cox regression in the problem of risky behavior parameter estimation based on the last episodes' data // St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 2021. vol. 14(4). pp. 202–217. DOI: 10.18721/JPM.14415.
13. Rahgozar M., Faghihzadeh S., Babaee Rouchi G., Peng Y. The power of testing a semi-parametric shared gamma frailty parameter in failure time data // Statistics in medicine. 2008. vol. 27. no. 21. pp. 4328–4339.
14. Balan T.A., Putter H. A tutorial on frailty models // Statistical methods in medical research. 2020. vol. 29. no. 11. pp. 3424–3454.
15. Czado C. Analyzing dependent data with vine copulas // Lecture Notes in Statistics, Springer. 2019. 242 p.
16. Nelsen R.B. An introduction to copulas (Springer Series in Statistics). Springer, 2006. 286 p.
17. Столярова В.Ф. Копулы и моделирование зависимости: косвенные оценки интенсивности рискованного поведения // Компьютерные инструменты в образовании. 2018. № 3. С. 22–37.
18. Kojadinovic I., Yan J. Comparison of three semiparametric methods for estimating dependence parameters in copula models // Insurance: Mathematics and Economics. 2010. vol. 47. no. 1. pp. 52–63.
19. Qian L., Zhao Y., Yang J., Li H., Wang H., Bai C. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes // Water Resources Management. 2022. vol. 36. no. 4. pp. 1141–1157.
20. Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В. Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. 2010. № 4. С. 30–38.
21. Suvorova A., Tulupyev A. Learning Bayesian network structure for risky behavior modelling // Proceedings of the Third International Scientific Conference "Intelligent

Information Technologies for Industry" (ITI'18). Springer International Publishing, 2019. pp. 58–65.

22. Суворова А.В., Тулупьев А.Л., Сироткин А.В. Байесовские сети доверия в задачах оценивания интенсивности рискованного поведения // Нечеткие системы и мягкие вычисления. 2014. Т. 9. № 2. С. 115–129.
23. Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей. СПб: СПбГУ, 2019. 399 с.
24. Koller D., Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009. 1230 p.
25. Langseth H., Nielsen T.D., Rumı R., Salmeron A. Mixtures of truncated basis functions // International Journal of Approximate Reasoning. 2012. vol. 53. no. 2. pp. 212–227.
26. Perez-Bernabe I., Maldonado A.D., Nielsen T.D., Salmeron A. Hybrid Bayesian Networks Using Mixtures of Truncated Basis Functions // R Journal. 2020. vol. 12. no. 2. pp. 321–341.
27. Scutari M., Denis J.-B. Bayesian Networks with Examples in R. 2nd edition. Chapman and Hall, Boca Raton. 2021. 274 p.
28. Czado C., Nagler T. Vine copula based modeling // Annual Review of Statistics and Its Application. 2022. vol. 9. no. 1. pp. 453–477.
29. Kolouri S., Kolouri S., Park S.R., Thorpe M., Slepcev D., Rohde G.K. Optimal mass transport: Signal processing and machine-learning applications // IEEE signal processing magazine. 2017. vol. 34. no. 4. pp. 43–59.
30. Hanea A.M., Hemming V., Nane G.F. Uncertainty quantification with experts: present status and research needs // Risk Analysis. 2022. vol. 42. no. 2. pp. 254–263.

Столярова Валерия Фуатовна — младший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: анализ данных, байесовские сети доверия, искусственный интеллект. Число научных публикаций — 40. vfs@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Тулупьева Татьяна Валентиновна — канд. психол. наук, доцент, старший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН); профессор кафедры, кафедра государственного и муниципального управления, Северо-Западный институт управления РАНХиГС. Область научных интересов: психология личности, искусственный интеллект, социальная инженерия, методы обработки данных. Число научных публикаций — 170. tvt@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Вяткин Артем Андреевич — младший научный сотрудник, лаборатория теоретических и междисциплинарных проблем информатики, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: байесовские сети доверия, искусственный интеллект. Число научных публикаций — 10. aav@dscs.pro; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-3337.

Поддержка исследований. Исследование выполнено в рамках инициативной НИР СЗИУ РАНХиГС при Президенте РФ, номер в системе ЕГИСУ НИОКТР 122112900066-6, а также в рамках проекта по государственному заданию СПБ ФИЦ РАН СПИИРАН № FFZF-2022-0003.

DOI 10.15622/ia.23.6.6

V. Stoliarova , T. Tulupyeva , A. Vyatkin

APPROACHES FOR BEHAVIOR INTENSITY ESTIMATION IN GROUPS OF HETEROGENEOUS INDIVIDUALS: PRECISION AND APPLICABILITY FOR DATA WITH UNCERTAINTY

Stoliarova V., Tulupyeva T., Vyatkin A. **Approaches for Behavior Intensity Estimation in Groups of Heterogeneous Individuals: Precision and Applicability for Data with Uncertainty.**

Abstract. In socially oriented areas, there arises the problem of assessing the cumulative characteristics of behavior, such as intensity, that are realized in groups of individuals. All individuals vary in their behavior and the available data is limited and may be associated with significant uncertainty: only a few episodes may be known and only a few individualsi the group may be observed. Mathematical models of behavior are used for estimation of key characteristics of the behavior. One of them is based on the gamma–Poisson point process, that reflects the heterogeneity of individuals in a form of a mixing distribution. This general model allows to formulate several methods of frequency estimation: the Cox regression, estimation of the copula parameter, and a posteriori inference in Bayesian belief networks. The aim of the paper is to assess their determine the precision of these methods based on the Kantorovich–Rubinstein distance between estimates and the true distribution of the desired parameter. The analysis of assumptions of those methods allows to formulate rules, that allow to chose the appropriate method in various sutuations of data availability. It has been shown that the copula-based approach provides the most accurate estimates and has the mild assumptions for the number of observed objects, but it cannot take into account external factors that may influence the behavior. Among methods that can take into account process covariants, estimates based on a posteriori inference in hybrid Bayesian belief networks have the highest precision. The paper considers a method for quantification of a hybrid BBNs with the approximation of mixtures of truncated exponents, that is data-demanding at the stage of calculating a priori estimates. However, it is noted that there are other approaches to setting hybrid BSDs in which a priori estimates can be set completely expertly.

Keywords: last episodes, risky behavior, Bayesian belief networks, hybrid bayesian belief networks, copula.

References

1. Chavez K., Palfai T.P., Cheng D.M., Blokhina E., Gnatienko N., Quinn E.K., Krupitsky E., Samet J.H. Hazardous Alcohol Use, Impulsivity, and HIV-Risk Behavior Among HIV-Positive Russian Patients With a History of Injection Drug Use. The American journal on addictions. 2021. vol. 30. no. 2. pp. 164–172.
2. Hendrieckx C., Ivory N., Singh H., Frier B.M., Speight J. Impact of severe hypoglycaemia on psychological outcomes in adults with type 2 diabetes: a systematic review. Diabetic Medicine. 2019. vol. 36. no. 9. pp. 1082–1091.
3. Paschenko A., Tulupyev A., Tulupyeva T., Krasnoselskikh T., Sokolovsky E. [Indirectassessment of the probability of HIV infection based on data on recent episodes of risky behavior]. Zdravoohranenie Rossijskoj Federacii – Healthcare of the Russian Federation. 2010. no. 2. pp. 32–35. (In Russ.).
4. Wojciechowski T.W. Major depressive disorder as a moderator of the relationship between heavy-episodic drinking and anxiety symptoms. Journal of mental health. 2023. pp. 1–8. DOI: 10.1080/09638237.2023.2245889.

5. Lewer D., Freer J., King E., Larney S., Degenhardt L., Tweed E.J., Hope V., Harris M., Millar T., Hayward A., Ciccarone D., Morley K. Frequency of health-care utilization by adults who use illicit drugs: a systematic review and meta-analysis. Addiction. 2020. vol. 115. no. 6. pp. 1011–1023.
6. Feldhege J., Moessner M., Bauer S. Who says what? Content and participation characteristics in an online depression community. Journal of Affective Disorders. 2020. vol. 263. pp. 521–527.
7. Jiotsa B., Naccache B., Duval M., Rocher B., Grall-Bronnec M. Social media use and body image disorders: Association between frequency of comparing one's own physical appearance to that of people being followed on social media and body dissatisfaction and drive for thinness. International journal of environmental research and public health. 2021. vol. 18. no. 6. DOI: 10.3390/ijerph18062880.
8. Oliseenko V., Khlobystova A., Korepanova A., Tulupyeva T. [Automatization of the assessment of the temperament of users of an online social networ]. Doklady Rossijskoj akademii nauk. Matematika, informatika, processy upravlenija – Reports of the Russian Academy of Sciences. Mathematics, computer science, management processes. 2023. vol. 514. no. 2. pp. 235–241. DOI: 10.31857/S2686954323601471. (In Russ.).
9. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users. Intelligent Distributed Computing XIII. Springer International Publishing, 2020. pp. 272–277.
10. Grandell J. Mixed Poisson Processes. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC. 1997. 280 p.
11. Cook R.J., Lawless J.F. The statistical analysis of recurrent events. Springer New York, 2007. 404 p.
12. Stoliarova V.F., Tulupyev A.L., Cox regression in the problem of risky behavior parameter estimation based on the last episodes' data. St. Petersburg Polytechnical State University Journal. Physics and Mathematics. 2021. vol. 14(4). pp. 202–217. DOI: 10.18721/JPM.14415.
13. Rahgozar M., Faghihzadeh S., Babaee Rouchi G., Peng Y. The power of testing a semi-parametric shared gamma frailty parameter in failure time data. Statistics in medicine. 2008. vol. 27. no. 21. pp. 4328–4339.
14. Balan T.A., Putter H. A tutorial on frailty models. Statistical methods in medical research. 2020. vol. 29. no. 11. pp. 3424–3454.
15. Czado C. Analyzing dependent data with vine copulas. Lecture Notes in Statistics, Springer. 2019. 242 p.
16. Nelsen R.B. An introduction to copulas (Springer Series in Statistics). Springer, 2006. 286 p.
17. Stoliarova V. Copula and dependency modeling: indirect estimates of the intensity of risky behavior. Komp'yuternye instrumenty v obrazovanii – Computer tools in education. 2018. no. 3. pp. 22–37. (In Russ.).
18. Kojadinovic I., Yan J. Comparison of three semiparametric methods for estimating dependence parameters in copula models. Insurance: Mathematics and Economics. 2010. vol. 47. no. 1. pp. 52–63.
19. Qian L., Zhao Y., Yang J., Li H., Wang H., Bai C. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes. Water Resources Management. 2022. vol. 36. no. 4. pp. 1141–1157.
20. Suvoriva A., Tulupyev A., Paschenko A., Tulupyeva T., Krasnoselskikh T. Analysis of granular data and knowledge in the problems of researching socially significant behaviors.

Komp'yuternye instrumenty v obrazovanii – Computer tools in education. 2010. no. 4. pp. 30–38. (In Russ.).
21. Suvorova A., Tulupyev A. Learning Bayesian network structure for risky behavior modelling. Proceedings of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18). Springer International Publishing, 2019. pp. 58–65.
22. Suvorova A., Tulupyev A., Sirotkin A. [Bayesian belief networks for the problems of assessing the intensity of risky behavior]. Nechetkie sistemy i myagkie vychisleniya – Fuzzy Systems and Soft Computing. 2014. vol. 9. no. 2. pp. 115–129. (In Russ.).
23. Tulupyev A.L., Nikolenko S.I., Sirotkin A.V. Osnovy teorii bajesovskih setej – Basics of Bayesian network theory. SPb: SPbGU, 2019. 399 p. (In Russ.).
24. Koller D., Friedman N. Probabilistic graphical models: principles and techniques. MIT press, 2009. 1230 p.
25. Langseth H., Nielsen T.D., Rumı R., Salmeron A. Mixtures of truncated basis functions. International Journal of Approximate Reasoning. 2012. vol. 53. no. 2. pp. 212–227.
26. Perez-Bernabe I., Maldonado A.D., Nielsen T.D., Salmeron A. Hybrid Bayesian Networks Using Mixtures of Truncated Basis Functions. R Journal. 2020. vol. 12. no. 2. pp. 321–341.
27. Scutari M., Denis J.-B. Bayesian Networks with Examples in R. 2nd edition. Chapman and Hall, Boca Raton. 2021. 274 p.
28. Czado C., Nagler T. Vine copula based modeling. Annual Review of Statistics and Its Application. 2022. vol. 9. no. 1. pp. 453–477.
29. Kolouri S., Kolouri S., Park S.R., Thorpe M., Slepcev D., Rohde G.K. Optimal mass transport: Signal processing and machine-learning applications. IEEE signal processing magazine. 2017. vol. 34. no. 4. pp. 43–59.
30. Hanea A.M., Hemming V., Nane G.F. Uncertainty quantification with experts: present status and research needs. Risk Analysis. 2022. vol. 42. no. 2. pp. 254–263.

Stoliarova Valerie — Junior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: data analysis, probabilistic graphical models. The number of publications — 40. vfs@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Tulupyeva Tatiana — Ph.D., Associate Professor, Senior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS); Professor of the department, Department of state and municipal management, North-West Institute of management (NWIM), the branch of the Russian Presidential Academy of National Economy and Public Administration. Research interests: personal psychology, artificial intelligence, social engineering, data analysis. The number of publications — 170. tvt@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Vyatkin Artyom — Junior research fellow, Laboratory of theoretical and interdisciplinary problems of computer science, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: data analysis, probabilistic graphical models. The number of publications — 10. aav@dscs.pro; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-3337.

Acknowledgements. The study was carried out within the framework of the initiative research NWIM branch RANEPA, № 122112900066-6 and within the framework of the project under the state order of the St. Petersburg Federal Research Center of the Russian Academy of Sciences SPIIRAN No. FFZF-2022-0003.

DOI 10.15622/ia.23.6.7

J. JACOB, K. KANNAN

ENHANCED MACHINE LEARNING FRAMEWORK FOR AUTONOMOUS DEPRESSION DETECTION USING MODWAVE CEPSTRAL FUSION AND STOCHASTIC EMBEDDING

Jacob J., Kannan K. **Enhanced Machine Learning Framework for Autonomous Depression Detection Using Modwave Cepstral Fusion and Stochastic Embedding.**

Abstract. Depression is a prevalent mental illness that requires autonomous detection systems due to its complexity. Existing machine learning techniques face challenges such as background noise sensitivity, slow adaptation speed, and imbalanced data. To address these limitations, this study proposes a novel ModWave Cepstral Fusion and Stochastic Embedding Framework for depression prediction. Then, the Gain Modulated Wavelet Technique removes background noise and normalises audio signals. Difficulties with generalisation, which results in a lack of interpretability, hinder extracting relevant characteristics from speech. To address these issues, an Auto Cepstral Fusion extracts relevant features from speech, capturing temporal and spectral characteristics caused by background voice. Feature selection becomes imperative when choosing relevant features for classification. Selecting irrelevant features can result in overfitting, the curse of dimensionality, and less robustness to noise. Hence, the Principal Stochastic Embedding technique handles high-dimensional data, minimising noise influence and dimensionality. Furthermore, the XGBoost classifier differentiates between depressed and non-depressed individuals. As a result, the proposed method uses the DAIC-WOZ dataset from USC for detecting depressions, achieving an accuracy of 97.02%, precision of 97.02%, recall of 97.02%, F1-score of 97.02%, RMSE of 2.00, and MAE of 0.9, making it a promising tool for autonomous depression detection.

Keywords: depression detection, machine learning, ModWave Cepstral Fusion, background noise, XGBoost classifier, DAIC-WOZ dataset, autonomous detection system, accuracy.

1. Introduction. Depression is the most significant reason for non-fatal health loss. In 2017, there were 322 million individuals worldwide who suffered from depression, according to the World Health Organization [1]. Depression can cause self-harm in addition to having a severe negative influence on one's family, career, and educational performance. Depression in adolescence is linked to mood disorders and severe mental illness in later life [2]. While depression is most commonly identified in those in their 30s and 40s, it can also be seen in older adults, youngsters, and those under stress in their relationships and academics [3], also frequently producing results comparable to those of major depression, minor depression, however less severe, adds significantly to the economic and social burden [4, 5]. As the most common mental illness in the world, MDD affects approximately 300 million individuals and is associated with significant financial burden and impairment [6, 7]. In Brazil, depression is the sixth most common health problem, with a lifetime prevalence of up to 16.8% and a prevalence rate of 5.8%. Depression symptoms include low

mood, irritation, anhedonia, fatigue, psychomotor slowness, cognitive impairment, and disturbances in internal systems. Early identification of depression symptoms, such as modified speech patterns, may facilitate immediate action and help avert the development of suicidal thoughts and poor social function [8].

A vital component of healthcare is mental health assessment, which enables early intervention and individualised therapy for people in psychological distress. Clinical examinations and subjective self-reporting are the foundations of conventional evaluation approaches, and they can be costly, time-consuming, or biased. However, new technological developments, especially in machine learning (ML), have opened the door for creative methods of mental health evaluation [9]. Using voice recordings in conjunction with ML techniques is one such potential method for identifying and tracking mental health issues. Tone, pitch, rhythm, articulation, and other aspects of voice carry much information that can reveal underlying emotional states and cognitive processes. Studies have demonstrated that people suffering from mental health conditions such as sadness, anxiety, and schizophrenia display unique patterns in their speech characteristics [10].

By analysing these fine-grained audio characteristics from speech recordings, ML techniques may detect patterns linked to particular mental health issues. These algorithms can acquire the ability to discriminate between normal and abnormal speech patterns with a high degree of accuracy by training on many annotated voice samples [11]. Furthermore, as time passes, ML algorithms can adjust and improve, constantly enhancing their prediction power. There are various benefits to combining ML and voice recording in mental health assessments. It offers a scalable, affordable, and non-invasive way to test people who could be at risk of mental health issues [12]. It also makes it possible to continuously track how well patients are doing and how they are responding to treatment, which makes tailored treatments easier to implement and enhances clinical results.

Common symptoms of depression include depressed emotions, loss of interest, mental slowness, and other symptoms. It is challenging to diagnose and has a protracted therapy cycle, a high incidence rate, and a sluggish onset [13, 14]. Psychotherapy and medication therapy are the primary forms of treatment. However, the diagnosis of depression has several deficiencies. First of all, depression is a prevalent mental illness, but many individuals avoid getting active therapy because they are embarrassed to admit they have it. Second, the widely utilised instrument for subjectively diagnosing depression is the Diagnostic and Statistical Manual of Mental

Disorders (DSM-5) [15, 16]. Misdiagnoses and missed diagnoses result from this. Thirdly, patients frequently lack the expertise required for self-assessment, and large-scale, low-cost depression screening instruments are absent [17]. As an outcome, many patients are unaware of their condition, which restricts their options for treatment. Finding an objective technique for quick screening and early warning for depression is therefore essential.

Several studies have shown an association between depression and an individual's behaviour. Research has shown that voice recordings of people can be a beneficial tool for characterising mental health and can offer significant insights into people's mental health. Furthermore, it has been proposed that precise results for depression prediction can be achieved by refining a combination of features. Since Mel-Frequency Cepstral Coefficients (MFCC) are trustworthy and effective even at low dimensionalities, they are the most often employed feature for audio signal processing. Several studies have demonstrated the effectiveness of an algorithm developed via ML techniques in detecting depression in voice samples. This study offers the following:

– With an autonomous system for detecting depression, the proposed ModWave Cepstral Fusion and Stochastic Embedding Framework addresses the complexity of conventional clinical diagnosis methods and offers an effective solution to the increasing number of depression cases worldwide.

– The proposed strategy, which introduces the Gain Modulated Wavelet Method, improves the quality of pre-processed data by efficiently removing background noise from audio recordings, normalising amplitude levels, and capturing both low and high-frequency information.

– The Auto Cepstral Fusion Feature extraction method is utilised to extract relevant features, thereby minimising the impact of noise, improving the robustness of the model, and capturing the temporal and spectral characteristics essential for depression prediction.

– Moreover, the Principal Stochastic Embedding method reduces dimensionality, minimises noise impact, and manages high-dimensional data, enhancing feature selection and classification precision.

– In addition to effectively predicting depression, the proposed approach thoroughly studies performance parameters like accuracy, precision, ROC, F1 score, recall, and sensitivity, offering valuable data for model comparison and evaluation.

The article will discuss recent studies on depression detection in machine learning, proposed methods and their explanations, the outcomes of the proposed work, and future directions with references.

2. Related Structure. Distinctive speech patterns such as lower articulation rate, pauses, slower speaking, lesser intensity, and unusual voice quality can be used to diagnose depression. Pitch, intensity, rhythm, speed, jitter, shimmer, energy distribution, and cepstral characteristics are examples of speech's prosodic, phonetic, and spectral aspects that must be taken into account to identify fluctuations in emotional state. Because jitter is sensitive to abrupt changes in speech, it is essential for identifying mood states. Cepstral coefficients – in particular, MFCC – have been well-researched for vocal analysis-based depression diagnoses that are well-suited for identifying depression speech.

The study by the authors in study [18] aimed to develop an ML tool for diagnosing depressive disorders. They used vocal feature extraction algorithms and ML classification techniques such as MLP, polynomial kernel SVM, polynomial kernel, normalised SVM PUK kernel, and random forest (RF) to extract vocal acoustic features from recordings. The results showed the tool's viability for cost-effective and non-invasive recognition and screening of MDDs, demonstrating its potential in diagnosing and screening these disorders. However, this technique lacks interpretability and also contains inconsistencies in voice, leading to misclassification of depressive disorders.

Paper [19] proposed a unique attention-based deep neural network that enables the merging of several modalities. This network is used to regress the depression level. This network has been trained using acoustic, text, and visual modalities. The regression process relies primarily on verbal input, which validates the therapist's experience. It can be challenging to combine text, graphics, and audio to estimate depression levels since integrating and synchronising multiple data sources is complicated and may require specialised technological knowledge and resources.

Study [20] extracted voice data features using Python programming and stored them in CSV files. For modelling, a database of 1479 voice feature samples was created. Utilising algorithmic selection and 10-fold cross-validation, a decision tree screening framework for depression was developed. Enhanced accuracy in forecasting was attained by the approach, enabling patients with depression to get early warning and care. It shows that clinical depression may be quickly identified and diagnosed using speech data. Depending on the complexity of this model, there is a risk of overfitting and limited generalizability.

An investigation on the creation of a supplementary tool for detecting depressive illnesses was carried out by the authors in paper [21], whereas 33 participants – 22 with a history of MDD and 11 healthy controls – were used to test automated classification algorithms and extract

voice acoustic characteristics. ML approaches and an approach for extracting vocal features were applied to the recordings. According to outcomes, random tree models with 100 trees outperformed other models in terms of categorisation, pointing to a non-invasive, low-cost technique for severe depressive illness detection and screening.

In study [22] suggested utilising multiple regression to predict the risk of depression using the context-DNN model. The expertise required to forecast circumstances and surroundings impacting depression while considering context information makes up the context of the suggested context-DNN. Every context data about depression predictor variables enters the DNN as an input, and each variable is used to predict the depression output of the DNN. Regression analysis was utilised to forecast the risk of depression for DNN connections to predict the possible context that may influence that risk. Due to their high learning capacity, DNNs may overfit, mainly when working with noisy or limited datasets.

To automatically identify depressed individuals on social media and provide an explanation for the model prediction, the authors in [23] suggested explainable Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN). They've considered user posts that had been enhanced with extra Twitter functionality. Specifically, the author computes the relevance of each tweet and word, encodes user posts using two levels of attention mechanisms applied at the tweet and word levels, and extracts semantic sequence features from user timelines (posts). The hierarchical attention approach was designed to identify patterns that provide interpretable outcomes. However, this model may have problems comprehending complicated models and raising privacy issues because the data it uses is sensitive.

This cross-sectional, descriptive-analytical study involved 205 pregnant Iranian patients under the care of Tabriz health centres. Cluster sampling was the sampling technique employed by the authors in paper [24]. Pregnant women completed the online Depression, Anxiety and Stress Scale-21 (DASS-21) and the sociodemographic characteristics questionnaire as part of the data-gathering process. The general linear model was employed to ascertain the components that were predictive of stress, anxiety, and depression. If a sample is not randomly selected or consists exclusively of people from a particular demographic, the study may be biased towards selection.

As a result, there were many restrictions on the method for identifying depression-related disorders. It must be interpretable and encounters consistent voice data, which could result in incorrect classifications. It takes specialised technological knowledge and resources

to integrate and synchronise text, video, and audio to determine depression levels. The existing works could have limited generalizability and run the risk of overfitting, especially with noisy or small datasets. Although it provides a low-cost, non-invasive method for detecting and screening severe mental disorders, there were several significant obstacles, including the model's complexity, potential biases in sample selection, and privacy issues because the data was sensitive.

3. Proposed Methodologies. ML algorithms have shown potential in detecting voice signal depression, potentially transforming mental disorder diagnosis. However, challenges remain, such as background noise sensitivity, limited adaptation speed, and low signal clarity due to class imbalances. It is challenging to reliably distinguish depression detection from speech signals due to these issues. Despite ongoing efforts to improve precision and accuracy, ML efficacy in mental health assessment remains limited by issues like interpretability, robustness, computational complexity, generalisation issues, overfitting, and dimensionality reduction. This study presents a novel ML paradigm for mental health depression detection, aiming to advance mental health diagnostics by enabling more precise and scalable detection. A rising number of individuals worldwide suffer from depression, a severe mental illness that affects people of any age. Traditional methods for diagnosing depression through mental health evaluations are complex and require machine learning techniques. However, limitations such as background noise sensitivity, less adaptation speed, and imbalanced data can impair the accuracy of existing machine learning systems. This study proposes a novel ML framework, ModWave Cepstral Fusion and Stochastic Embedding Framework, to predict depression. To overcome these challenges, a Gain Modulated Wavelet Technique is employed to remove background noise from audio recordings, capturing low- and high-frequency information. The next step is feature extraction, which reduces noise impact and improves model robustness. The Auto Cepstral Fusion Feature extraction technique is introduced to capture temporal and spectral characteristics caused by background voice. Feature selection is crucial for classification, as selecting irrelevant features can lead to overfitting, the curse of dimensionality, less robustness to noise, and low interpretability. The Principal Stochastic Embedding technique handles high-dimensional data, minimises noise influence, and enhances model performance. Classification is performed using the XGBoost classifier to determine if a person is depressed. Figure 1 shows the proposed workflow diagram comprising pre-processing, feature extraction, feature selection, and classification procedures.

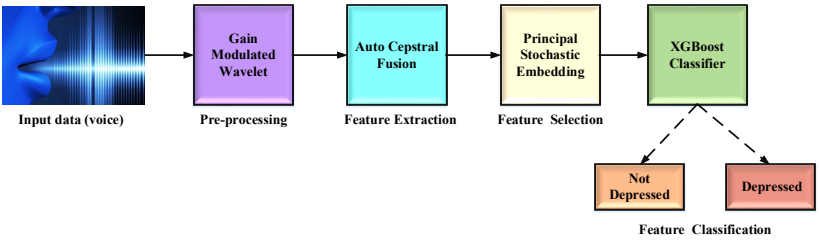


Fig. 1. Proposed Workflow diagram

3.1. Pre-processing based on Gain Modulated Wavelet Technique. Gain Modulated Wavelet (GMW) Technique removes background noise from audio recordings, capturing both low- and high-frequency information in voice signals and normalising audio signals to ensure consistent amplitude levels across recordings. Pre-processing of audio signals typically involves several steps aimed at enhancing the signal's quality or extracting useful information from it. To remove noise from audio signals, the GMW technique strengthens the adaptive gain model and discrete wavelet transform methods. Pre-processing audio signals with adaptive gain control and discrete wavelet transform can be effective for several functions, including audio denoising and capturing information at low and high frequencies.

Decompose the audio signal $x(n)$ into wavelet coefficients using discrete wavelet transform (DWT), which is a powerful tool for time-frequency analysis. The decomposition of the audio signal into wavelet coefficient using DWT is determined in equation (1):

$$x(n) = \sum_{k=0}^{N-1} c_{J,k}\, \phi_{J,k}(n) + \sum_{j=1}^{J} \sum_{k=0}^{N/2^{j}-1} d_{j,k}\, \psi_{j,k}(n), \quad (1)$$

where $c_{J,k}$ are the approximation coefficients at scale $J, d_{j,k}$ are the detail coefficients at scale j, and $\phi J,k$ and $\psi j,k$ are the scaling and wavelet functions, correspondingly.

Estimate the noise level σ_j in each detail coefficient sub-band employing robust methods, including median absolute deviation (MAD) or local variance estimation. The estimation of noise level is measured in equation (2):

$$\sigma_j = \mathrm{MAD}(d_{j,k}) / 0.6745, \quad (2)$$

where σ_j represents the estimated noise level in the j^{th} detail coefficient sub-band. $MAD(d_{j,k})$ refers to the mean absolute deviation of wavelet

coefficients in j^{th} sub-band. 0.6745 is a constant scaling factor used to scale MAD to estimate standard deviation for normally distributed data. Combining everything, the formula determines the noise level (σ_j). The $MAD(d_{j,k})$ of the wavelet coefficients in a given sub-band is obtained by dividing it by a scaling factor (0.6745).

Apply adaptive gain control to the detail coefficients using the estimated noise levels. The gain factor g_j for each sub-band is calculated in equation (3):

$$\mathrm{g_j} = \frac{\sigma_{\mathrm{target}}}{\sigma_{\mathrm{j}}}, \tag{3}$$

where σ_{target} is the target noise level. Normalise the detail coefficients by multiplying them with the respective gain factors g_j. The normalised coefficients are determined in equation (4):

$$\mathrm{d'_{j,k}} = \mathrm{g_j}.\,\mathrm{d_{j,k}}, \tag{4}$$

where $d'_{j,k}$ are the modified detail coefficients after gaining control. Reconstruct the denoised signal $x'(n)$ by applying the inverse discrete wavelet transform to the modified coefficients. The reconstruction of the denoised signal is calculated in equation (5):

$$\mathrm{x'(n)} = \sum_{\mathrm{k=0}}^{\mathrm{N-1}} \mathrm{c'_{J,k}}\,\phi_{\mathrm{J,k}}(\mathrm{n}) + \sum_{\mathrm{j=1}}^{\mathrm{J}} \sum_{\mathrm{k=0}}^{\mathrm{N}/2^{\mathrm{j}}-1} \mathrm{d'_{j,k}}\,\psi_{\mathrm{j,k}}(\mathrm{n}). \tag{5}$$

Adaptive gain control and discrete wavelet transform are combined in this hybrid method to efficiently reduce noise in the audio stream while maintaining significant signal characteristics. The raw data is normalised to eliminate background noise from audio recordings, capturing both low- and high-frequency information present in voice signals and normalising audio signals to ensure consistent amplitude levels across recordings using the combined power of adaptive gain control and discrete wavelet transform. The next stage is to extract features from pre-processed data. The dataset was balanced using the SMOTE approach, which comes after the noise removal procedure. The Synthetic Minority Over-sampling Technique (SMOTE) is a machine learning algorithm that addresses class imbalance, where a minority class is underrepresented in a dataset, leading to biased models. SMOTE balances class distribution, making the model more robust and less biased towards the majority class. Although its effectiveness

depends on the dataset and problem, it is a valuable tool for handling imbalanced datasets by generating synthetic samples for the minority class, improving machine learning model performance, especially in situations where the minority class is of significant interest. The process of SMOTE algorithm could be divided into several steps:

1. To create additional samples, choose one minority sample and generate it as x_i.

2. Determines k closest neighbors x_i. Make a random selection from them and mark it with x_j.

3. To create a new sample of x_{new}. Equation (6) uses θ, a random number between 0 and 1.

$$x_{new} = x_i + \theta * (x_j - x_i). \quad (6)$$

4. Repeat steps 2 and 3 a total of round $(N/100)$ times to create round $(N/100)$ minority samples.

5. Apply the aforementioned process to every minority sample $(i = 1,2,\dots,T)$.

In imbalance situations, the SMOTE method produces minority samples to improve classifier performance and balance datasets. All samples are treated equally, though, thus it might miss samples that could be mistakenly labelled. Proposed algorithms for over-sampling aim to increase the accuracy of minority samples by giving greater weights to samples that are prone to misclassification. In unbalanced datasets, this method guarantees that minority samples are given greater weight than majority samples.

3.2. Feature extraction based on Auto Cepstral Fusion technique. In audio signal processing, extracting relevant features from pre-processed data is crucial for various applications. However, difficulties arise with generalisation and interpretation due to the raw nature of voice signals, leading to poor performance in feature extraction. A technique called Auto Cepstral Fusion feature extraction is introduced to address these challenges. This method combines Autocorrelation and Mel Frequency Cepstral Coefficients (MFCCs) to enhance feature extraction capabilities. Autocorrelation and MFCCs are combined to capture spectral and temporal information from audio signals, making the technique versatile and practical for various audio analysis tasks. The process begins with pre-processed audio signals. MFCC extraction is applied to capture the spectral envelope, while autocorrelation extracts the audio's pitch period and harmonic structure. Combining these techniques allows for a comprehensive

understanding of audio signals, improving interpretability and performance in feature extraction tasks.

3.2.1. Mel Frequency Cepstral Coefficients (MFCCs). Feature extraction is the technique of considering a stationary speech segment that is small enough while identifying and computing a collection of features for every short time frame of the input speech signals to provide meaningful modelling. Since the MFCC method's computation depends on short-time power, features are extracted in this study utilising the mel-frequency cepstral coefficient. The spectrum obtained from the vocal cords of humans further maps the known fluctuation of critical bandwidth frequencies of the human ear using two filters to capture the essential components of speech: a logarithmic filter at high frequencies above 1 kHz and a linear filter at low frequencies below 1 kHz. Figure 2 illustrates the MFCC feature extraction process. The MFCC includes some extraction process which follows.

Pre-emphasis. It needs to go through a filter to make up for the high-frequency part muted through the human sound-generating process. The high pass filter is applied to the voice signal in equation (7):

$$x_1(n) = x(n) - \alpha * x(n-1), \tag{7}$$

where $x_1(n)$ represents the output signal, $x(n)$ and $x(n-1)$ signifies present and past signal individually. The value α lies between 0.9 to 1.

Frame Blocking. The continuous speech signal is split into N sample-sized frames, with N-M samples overlapping and M samples (M<N) separating neighbouring frames. This procedure keeps going till the signal is divided into smaller frames.

Windowing: the windowing process involves tapering the signal to zero at the beginning and end of every frame to reduce spectral distortion. After multiplying the signal $x(n)$ by a window $w(n)$ at time n, the extracted signal is obtained by equation (8):

$$y_2(n) = x(n) * w(n), \quad 0 \leq n \leq N-1, \tag{8}$$

where N is the number of samples in every frame. Since the Hamming window sinks the sidelobe level of window transfer while reducing the frequency resolution of spectral analysis, it is used in this case; the spectral analysis for reducing the frequency resolution is determined in equation (9):

$$w(n) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right], \quad 0 \leq n \leq N-1. \tag{9}$$

Fast Fourier Transform. Transfers N frequency domain samples to the time domain. The Discrete Fourier Transform (DFT), which depends on a collection of N Samples (yn), is developed using the widely used FFT approach. The estimation of the DFT process is determined in equation (10):

$$Y_n = \sum_{n=0}^{N-1} y_n e^{-j2\pi kn/N}, \; k = 0,1,2 \ldots \ldots, N-1. \quad (10)$$

A spectrum or periodogram is a concept used to describe the outcome of the FFT process.

Mel-frequency wrapping. Mel frequency depends primarily on research on how humans perceive frequency. All frequency bands exhibit varying degrees of sensitivity in human hearing. It becomes less responsive to increased frequencies over 1000 Hz. Mel-frequency, defined as linear frequency spacing below 1 kHz, is the voice signal. The estimation of the Mel-frequency wrapping process is measured in equation (11):

$$Mel(f) = 2595 * \log_{10}(1 + f/700). \quad (11)$$

Cepstrum. In this final step of the MFCC procedure, the log mel spectrum is transformed into the time domain. Since DCT's results include significant quantities of energy, DCT typically conducts this conversion. The DCT output is expressed as MFCC and is represented in equation (12):

$$C[n] = \sum_{n=0}^{N-1} \log \left| \sum_{n=0}^{N-1} x(n) \exp\left(\frac{-j2\pi kn}{N}\right) \right| \exp\left(\frac{j2\pi kn}{N}\right), \quad (12)$$

where n=0, 1,2,……N-1. $C[n]$ means MFCC, and twelve cepstral coefficients are retrieved from every frame, where n is the number of coefficients (n=12).

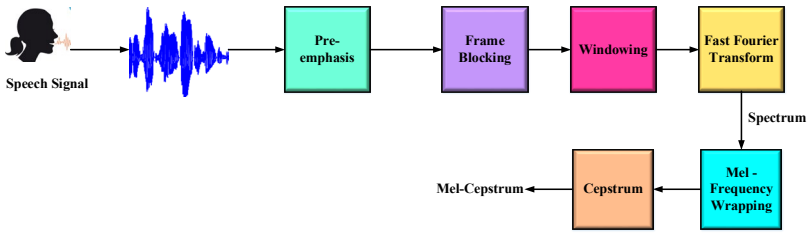


Fig. 2. MFCC feature extraction process

3.2.2. Auto Correlation. The autocorrelation function compares the similarity between the delayed and original signals to determine how self-similar a signal is in the temporal domain. A strong positive association is indicated by an autocorrelation value of +1, a negative association by -1, and no association by 0. Because the signal has a perfect correlation with itself, the autocorrelation at lag zero is always 1. Autocorrelation is instrumental in capturing periodic and repetitive patterns in speech signals. It's computed by correlating the signal with itself at various time lags, revealing crucial speech characteristics like formants and pitch. Calculate the autocorrelation function for each speech signal frame to better understand its self-similarity. The autocorrelation function is defined in equation (13):

$$\mathrm{R(k)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{x(n).x(n+k)}, \tag{13}$$

where N is the frame length, $x(n)$ represents the signal at the time index n, k is the lag at which autocorrelation is computed, and $R(k)$ is the autocorrelation value at lag k. The autocorrelation function captures the self-similarity of the signal, highlighting periodic components. Autocorrelation values may vary depending on the amplitude and energy of the signal. Normalisation helps make the feature more robust and invariant to changes in amplitude. Each autocorrelation value is divided by the autocorrelation at lag 0 to normalise it. The autocorrelation process for normalising the value is measured in equation (14):

$$\mathrm{R'(k)} = \frac{\mathrm{R(k)}}{\mathrm{R(0)}}. \tag{14}$$

From the computed autocorrelation function, you can extract various features that are useful for speech recognition and periodicity of the signal, such as:

– Pitch period: the pitch period of the signal is frequently correlated with the lag corresponding to the first peak following lag 0.

– Harmonic Structure: the autocorrelation function's regularly spaced peaks can indicate that the signal has harmonic components.

– Envelope Information: the signal envelope can be obtained from the decay rate of the autocorrelation values. In audio and voice processing applications, autocorrelation facilitates the extraction of significant signal features, making tasks like pitch estimation, harmonic analysis, and envelope identification easier. While the autocorrelation

features offer more details about the temporal and periodic patterns, the MFCC features capture the speech's spectrum qualities. Auto Cepstral Fusion, a Feature extraction technique, extracts relevant features, reducing the impact of noise, improving the robustness of the model, and aiming to capture temporal and spectral characteristics caused by background voice. The proposed strategy provides feature selection to minimise dimensionality based on the particular application and requirements. This is described in more detail in the steps that follow and can aid in identifying the most valuable features while reducing computing complexity.

3.3. Feature Selection based on the Principal Stochastic Embedding technique. Selecting the most pertinent characteristics is the next step after feature extraction. The principal stochastic embedding technique, which supports t-distributed Stochastic Neighbor Embedding (t-SNE) along with Principal Component Analysis (PCA), is used to carry out this procedure. This combination allows for capturing global and local structures identified in voice signal data, making it possible to visualise complex relationships more thoroughly. PCA is good at capturing global structures, which helps it discover broad patterns and trends within the data. Still, t-SNE focuses on maintaining local structures, which allows it to capture complex interactions among neighbouring data points. When these two approaches work together, complicated relationships in the data can be shown more effectively, leading to a more informed feature selection technique.

3.3.1. t-SNE algorithm. The essential elements strongly associated with the target characteristic are selected by applying dimensional reduction techniques. Significant and highly representative features are collected to achieve high accuracy. Decreasing the amount of variables in a dataset is known as dimensional reduction. A proposed technique for reducing the dimensionality of nonlinear data is to drop it from a high-dimensional space into a low-dimensional one using the t-SNE technique. The method focuses on the variance of neighbourhood points and data inclusion in a low space, producing random, unconfirmed probability. It assigns comparable traits with greater probability and dissimilar characteristics to lower probability when distributing pairs of X_i and X_j. The pairwise similarity in the high-dimensional data space is determined in equation (15), and the data points representation by t-SNE in a low-dimensional space is demonstrated in equation (16). Equation (17) illustrates how the technique iteratively operates the same probability distribution across a smaller space to show data points in a low-dimensional space and lower the Kulback-Leibler (KL) variance. The probability distribution with low KL variance is determined in equation (17):

$$P(X_i/X_j) = \frac{S(X_i\ ,\ X_j)}{\sum_{m\neq i}^{N} S(X_i,\ X_m)}, \tag{15}$$

$$Q(Y_i/Y_j) = \frac{S(Y_i,\ \ Y_j)}{\sum_{m\neq i}^{N} S\ (Y_i,\ \ Y_m)}, \tag{16}$$

$$KL = \sum_i \sum_j P(X_i,\qquad X_j) \log \frac{P(X_i/X_j)}{Q(Y_i/Y_j)}, \tag{17}$$

where $P(X_i/X_j)$ high-dimensional data is space and X_i/X_j are pairs in the P space; Y_i/Y_j is low-dimensional data space, and Y_i/Y_j are pairs in the Q space (Figure 3).

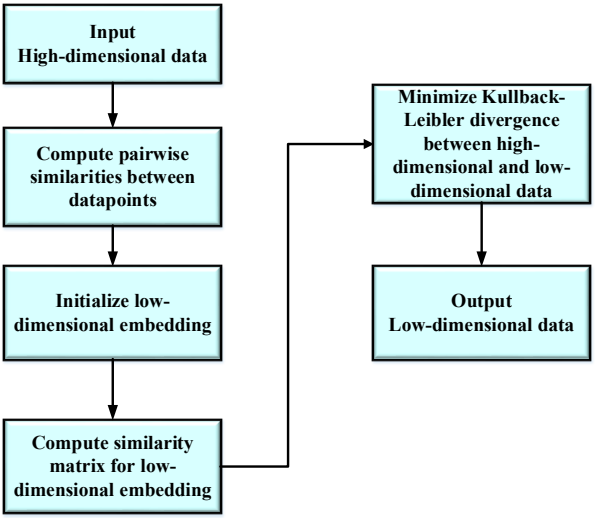


Fig. 3. Flowchart for t-SNE algorithm

3.3.2. PCA Algorithm. Principal components of PCA, an unsupervised statistical technique, are linearly uncorrelated when correlated features are converted. Normalising the dataset uses mathematical ideas like variance, covariance, eigenvalues, and eigenvectors. Correlation is the connection between two characteristics, whereas dimensions are the amount of features in the collection. To ensure there are high-variance features, the method divides the individual value by the standard deviation of all

features. The Z covariance matrix includes the variance among the two feature pairs, and eigenvectors represent high-variance information axes. In the P matrix, the technique places eigenvalues and eigenvectors in descending order. The Z covariance matrix is multiplied by the P matrix to generate new features. Important and pertinent characteristics are kept, while less important ones are eliminated to produce a new dataset. The dimensionality reduction feature selection technique aids in reducing features while preserving the most significant amount of relevant information. PCA identifies the principle components that primarily explain the variation in the data, whereas t-SNE produces a low-dimensional representation that preserves the local structure. The hybrid strategy lowers the risk of overfitting and improves model generalisation by combining both techniques to achieve more effective dimensionality reduction. This optimised feature selection process produces a subset of characteristics that are very discriminative and informative of the underlying structure in the data. Enhanced visualisation, complete data representation, improved dimensionality reduction, optimal feature selection, and increased performance in machine learning tasks are just a few benefits of the hybrid feature selection approach utilising principle stochastic embedding. Figure 4 shows the flowchart for the PCA algorithm. The next step involves a classification process, which classifies depressed patients.

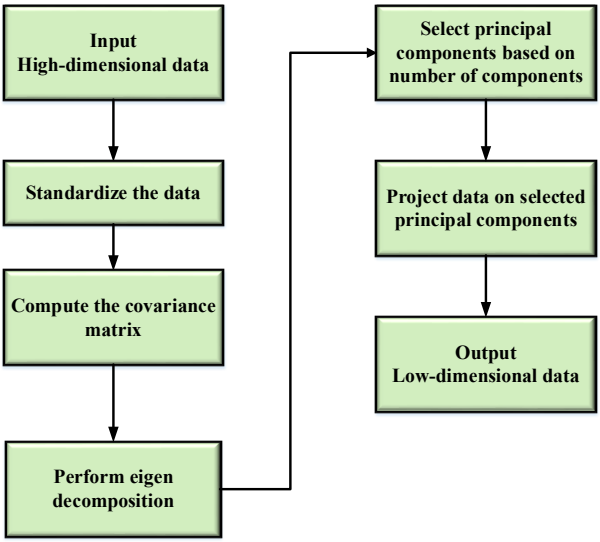


Fig. 4. Flowchart for PCA

3.4. XGBoost Classifier. Using the XGBoost classifier, also known as the Extremely Gradient Boosted Decision Tree, can effectively segment data into smaller subsets, leading to remarkable accuracy across many Natural Language Processing (NLP) applications and ML models. This classifier boasts several advantages, including scalability, parallelizability, and swift execution times, making it a preferred choice in various settings. Moreover, it is a regularised model in which formalisation helps to prevent overfitting, thereby enhancing performance compared to other algorithms. Utilising the XGBoost classifier, the proposed technique demonstrates superior performance in identifying depressed individuals compared to existing approaches. Through its expertise in handling intricate relationships within data and its robustness against overfitting, XGBoost ensures high accuracy and reliability in predicting outcomes, thereby offering a promising avenue for advancing depression detection. The XGBoost workflow schematic is shown in Figure 5. The blue-coloured zone represents the training and testing data. The boxes inside the dashed lines represent the testing and training procedures, where GBM stands for gradient boosting machine and T is for tree. The results obtained from XGBoost from the dashed box are displayed in the two oval boxes on the right.

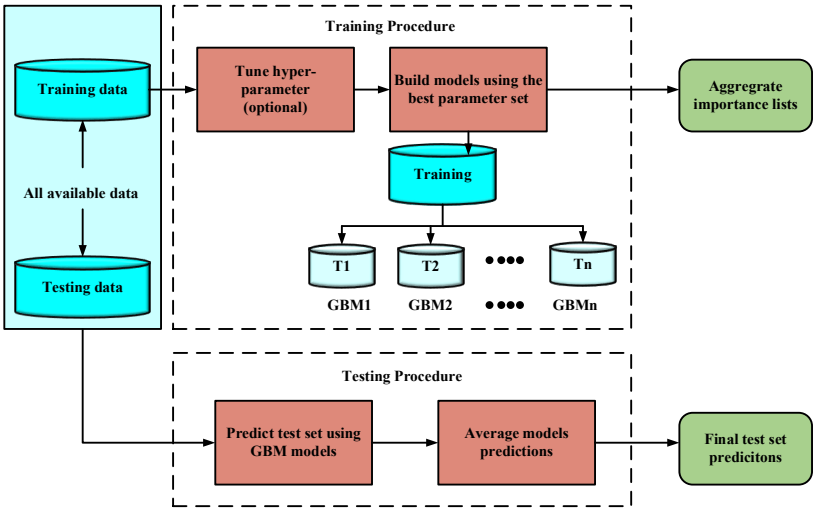


Fig. 5. XGBoost classifier

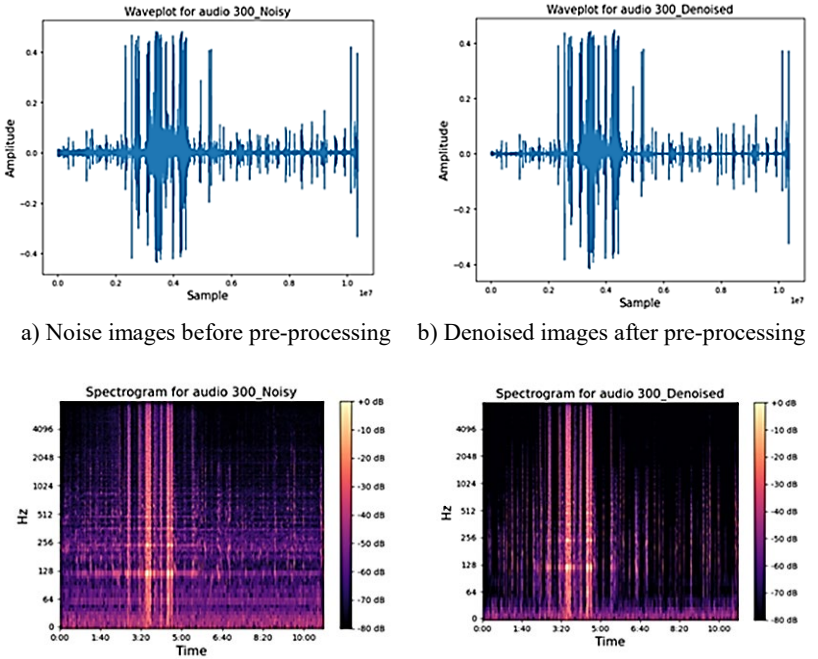
4. Results and Discussion. The Distress Analysis Interview Corpus: Wizard of Oz (DAIC-WOZ) dataset is employed in this research [25]. Based on feature vectors, data was chosen and split into two categories:

80% for training and 20% for testing. The dataset comprises text, audio, and video exchanges among individuals and an automated interviewer. The interview questions are based on the physical manifestations of depression. Two versions of this dataset have been released up to this point; the expanded DAIC-WOZ is utilised in this study. There are 189 sessions in the sample (102 men and 87 women), of which 133 are not depressed, and 56 are depressed. The dataset is split into 80% and 20% for training and testing. The outcomes achieved are presented in the subsections that follow. The Python tool, Windows 7 (64-bit) OS, and Intel Premium CPU with 8GB RAM are used to carry out this proposed work. In conclusion, it compared previous approaches and the proposed system. This section discusses how well the approach we propose works for identifying depressed patients.

4.1. Performance analysis. The amplitude of the audio signal is shown in the waveplot graph. Unequal or fluctuating amplitude in a noisy audio waveplot indicates the existence of background noise. The noisy waveplot implies that the audio signal has been affected by unwanted noise, which might affect the clarity of the audio. Figure 6(a) shows that the x-axis depicts the sample, and the y-axis is the audio signal's amplitude. The audio signal amplitude following noise reduction processing is seen in the denoised waveplot: a less amplitude fluctuating, smoother waveplot than the noisy version. The denoised waveplot shows that the noise reduction procedure has successfully eliminated or reduced background noise, producing a more precise and cleaner audio stream. Figure 6(b) shows that the x-axis depicts the sample, and the y-axis is the audio signal's amplitude.

A spectrogram shows the audio signal's frequency content with time. The presence of noise is shown in spectrograms of noisy audio as extra energy in different frequency bands, which frequently take the form of irregular patterns or streaks. A noisy spectrogram may mask or distort the properties of the underlying signal by highlighting spectral contamination carried on by background noise. Figure 6(c) shows that the y-axis indicates hz, and the x-axis represents time. The frequency content of the audio signal is seen on the denoised spectrogram following noise reduction. Reduced energy in background noise-corresponding frequency regions improves the visibility of signal characteristics and produces more apparent spectral patterns. A denoised spectrogram shows how noise reduction may improve the signal-to-noise ratio, which makes it easier to identify and analyse relevant audio properties. Figure 6(d) shows that the y-axis indicates Hz, and the x-axis represents time in seconds. With the wave plot showing changes in amplitude and the spectrogram indicating changes in frequency content, both offer useful information on how noise reduction affects audio

signals. These graphics help evaluate how well noise reduction methods work and how they affect the audio signal's overall quality.

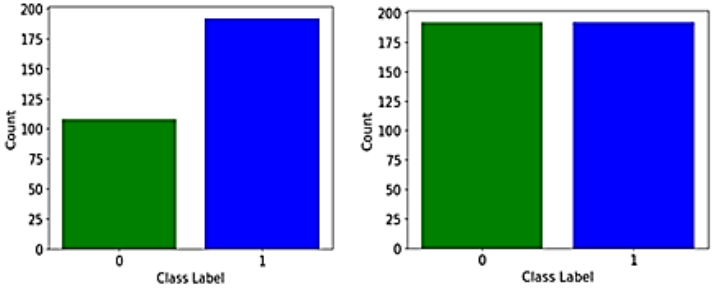


a) Noise images before pre-processing b) Denoised images after pre-processing

c) Noisy spectrogram audio d) Denoised spectrogram audio

Fig. 6. Wave plot and Spectrogram visualisation

4.2. Imbalanced Datasets. Figure 7 below shows the dataset modelling (before and after balancing the dataset). There are two types of class labels: 0 is represented as depressed, and 1 is represented as not depressed. The dataset utilised in this study comprises individuals categorised as depressed and not depressed, but it exhibits an inherent imbalance and is contaminated with background noise. Consequently, the pre-processing methodology outlined in this study addresses these challenges. Following pre-processing, the dataset achieves a balance between the depressed and not-depressed categories while also effectively eliminating background noise. Furthermore, this pre-processing method ensures the capture of low and high-level features within the dataset. Here, we used the SMOTE technique to balance the dataset, which was performed after the noise removal process.



a) Imbalanced dataset b) Balanced dataset

Fig. 7. Dataset modelling

4.3. Confusion matrix. One kind of performance metric employed in ML and classification to evaluate a model's ability to identify depression is a confusion matrix. It offers a summary of the variations among predicted labels and the actual ground truth labels so that the model's accuracy may be evaluated. The four groups in the confusion matrix are TP, FP, FN, and TN, where 0 denotes people who are depressed, and 1 represents those who are not depressed. Several performance measures, including accuracy, precision, recall, and F1-score, may be computed using the confusion matrix to assess how well the framework identifies depression. Figure 8 shows the confusion matrix for the proposed work.

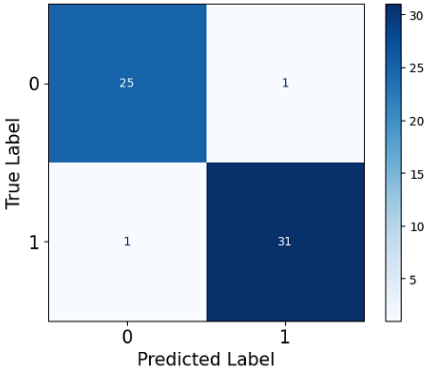


Fig. 8. Confusion matrix

4.4. ROC curve. A graphical depiction of a binary classification model's accuracy across various threshold values is called a Receiver Operating Characteristic (ROC) curve. At various threshold values, it shows the true positive rate (TPR) versus the false positive rate (FPR). The ROC

curve, shown in Figure 9, depicts the false positive rate (specificity) on the x-axis and the true positive rate (sensitivity) on the y-axis. An area under the curve (AUC) of 0.97 on an ROC curve for detecting depression denotes good discriminating power in differentiating between those who are depressed and those who are not. A high likelihood of ranking a randomly selected sad person higher than a randomly chosen non-depressed person is indicated by an AUC of 0.97. This suggests that the model makes few prediction errors, achieving high sensitivity while maintaining low false positive rates across various threshold settings.

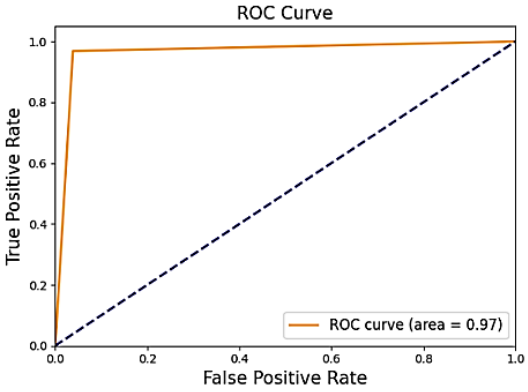


Fig. 9. ROC curve

4.5. Comparative analysis. The depression prediction framework was evaluated with several performance metrics: root mean square error (RMSE), mean absolute error (MAE), accuracy, precision, recall and F1-score. The performance metrics are compared with several existing works such as Deep Convolutional Neural Network-Deep Neural Network ((DCNN-DNN) [26], Deep Convolutional Generative Adversarial Network (DCGAN) [27], Transformer Encoder + Convolutional Neural Network (TE+ CNN) [28], Bidirectional-Long Short term Memory + Attention (Bi-LSTM + Attention) [29], Graph Convolutional Neural Network (GCNN) [30], Convolutional Neural Network (CNN) [31], Gated Recurrent Unit (GRU) [31], Bimodal Attention-GRU (BiAtt-GRU) [31], Two-dimensional CNN-LSTM (2D-CNN-LSTM) [32], Decision Tree (DT) [32], Deep AudioNet (DAN) [32], Transformer-CNN-CNN (TCC) – Softmax [32], Unimodal Ensemble (UE) [33], Multimodal + Selective dropout + Transfer Learning (MM + SD + TL) [33], Multimodal + Selective dropout-Normalization-Attention + Transfer Learning + Spectral-Normalized Neural Gaussian Process (MM+ SD-Norm-Att + TL + SNGP) [33].

The Root Mean Square Error (RMSE) is a frequently employed metric for assessing a model's or prediction's accuracy. The definition of this term is the square root of the average of the squared differences between the actual and predicted values. Mathematically, RMSE is represented in equation (18):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{18}$$

The accuracy of a model's predictions is gauged by the Mean Absolute Error (MAE) metric. The mean of the absolute differences between the actual and predicted values is how it is defined. Without considering direction, the mean absolute error (MAE) quantifies the average magnitude of mistakes in a set of predictions. A lower MAE value denotes a more precise model. The MAE is determined in equation (19):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{19}$$

where $\hat{y}_i$, y_i and n display the actual severity score, the predicted score generated from the model, and the amount of test data corresponding to it. By using equations (18) and (19), the RMSE and MAE values are calculated. The proposed method performs significantly fewer errors than other existing techniques. Figure 10 compares RMSE and MAE graphs for the proposed model with several existing works. Table 1 shows the comparison values for the RMSE and MAE metrics.

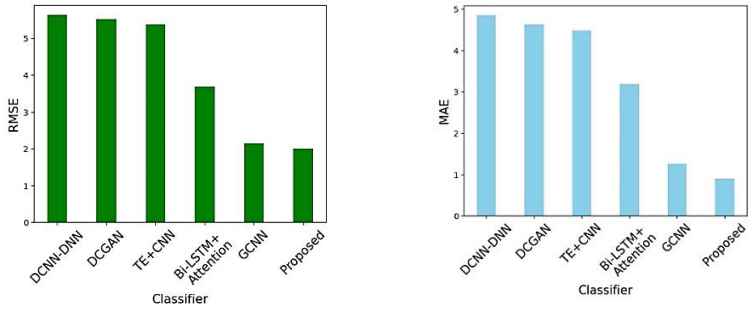


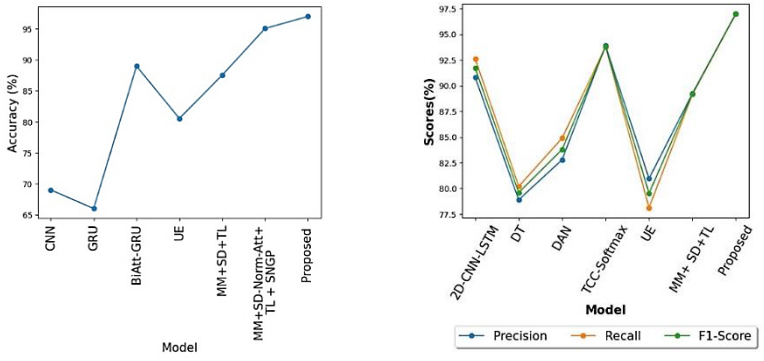
Fig. 10. Comparison graph for: a) RMSE; and b) MAE

Table 1. Comparison of error metrics

Method	RMSE	MAE
DCNN-DNN [26]	5.63	4.85
DCGAN [27]	5.52	4.63
TE+ CNN [28]	5.37	4.48
Bi-LSTM + Attention [29]	4.76	3.61
GCNN [30]	2.15	1.25
Proposed	2.00	0.90

The efficiency of the suggested strategy is demonstrated by comparing approaches for predicting depressive illnesses based on their Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). With an MAE of 1.25 and the lowest RMSE of 2.15, the GCNN demonstrates excellent prediction accuracy. The suggested approach, on the other hand, outperforms the others in terms of predicting the severity of depression, with an even lower RMSE of 2.00 and MAE of 0.90. While models such as the Bi-LSTM + Attention demonstrate competitive performance (RMSE of 4.76 and MAE of 3.61), they are less effective than the proposed technique. The Transformer Encoder + CNN and DCGAN approaches show slightly higher prediction errors, with corresponding MAE values of 4.48 and 4.63 and RMSE values of 5.37 and 5.52, respectively. The suggested approach stands out for its precision and accuracy in estimating the severity of depressive disorder, providing encouraging developments in this area of study.

Evaluation metrics. Accuracy, precision, recall, and F1 score were used as performance indicators for this study. These metrics finally demonstrate the proposed technique's performance reliability. Figure 11 below displays the comparison graph for the proposed work.



a) b)

Fig. 11. Comparison graph for performance metrics: a) Accuracy, b) Precision, Recall and F1-score

The work proposed provides all of the current methods, and Tables 2 and 3 display the comparative analysis of the performance metrics for the proposed model with several existing works.

Table 2. Comparison of Accuracy metrics

Model	Accuracy (%)
CNN [31]	69.00
GRU [31]	66.00
BiAtt-GRU [31]	89.00
UE [33]	80.54
MM + SD + TL [33]	87.55
MM + SD-Norm-Att + TL + SNGP [33]	95.07
Proposed	97.02

The suggested model outperforms multiple existing models in detecting depression-related conditions, with the maximum accuracy of 97.02%. By contrast, the accuracy of the GRU model is 66%, the CNN model is 69%, and the BiAttention-GRU model is 89%. While the Multimodal + SD + transfer learning model achieves 87.55% accuracy, the Unimodal Ensemble model only manages 80.54%. 95.07% is achieved by the Multimodal + SD-Norm-Att + transfer learning + SNGP model, demonstrating a notable improvement with the suggested approach.

Table 3. Comparison of Precision, Recall and F1-score

Model	Precision (%)	Recall (%)	F1-score (%)
2D-CNN-LSTM [32]	90.80	92.60	91.70
DT [32]	78.90	80.20	79.60
DAN [32]	82.80	84.90	83.80
TCC-Softmax [32]	93.90	93.80	93.80
UE [33]	80.96	78.13	79.52
MM + SD + TL [33]	89.26	89.17	89.21
Proposed	97.02	97.02	97.02

A high degree of accuracy and consistency in its predictions is indicated by the suggested model's precision, recall, and F1-score, which are all at 97.02%, suggesting outstanding performance in diagnosing depression. It performs better than a number of other models, including the TCC-Softmax model, which attains a precision, recall, and F1-score of 93.8%, and the 2D-CNN-LSTM, which has a precision of 90.8%, recall of 92.6%, and F1-score of 91.7%. Some models perform better than others, such as the Unimodal Ensemble and Deep AudioNet, which scored 79.52%

and 83.8%, respectively, on the F1 score. Furthermore, the F1-score for the Multimodal + SD + transfer learning model is 89.21%.

This research demonstrates the superiority of the novel method in terms of depression detection, with a low MAE of 0.9 and an exceptional accuracy of 97.02%. In terms of precision and forecast accuracy, this strategy outperforms earlier ones. The study also emphasises the significance of accuracy and F1-score, standard metrics in evaluation processes, in determining the model's efficiency. The proposed method is a potential growth in effective depression detection methods as it outperforms currently available methods and produces reliable findings regarding the accuracy and F1 score.

5. Conclusion. In conclusion, the high incidence of depression necessitates the development of autonomous detection systems, given the complexities associated with traditional clinical diagnosis methods. Existing ML techniques for depression detection encounter challenges such as sensitivity to background noise, slow adaptation speed, and imbalanced data, which can compromise accuracy. To overcome these limitations, this study introduces a novel ModWave Cepstral Fusion and Stochastic Embedding Framework for depression prediction. They address issues like background noise in audio signals and low amplitude levels during pre-processing by employing the Gain Modulated Wavelet Technique. This technique removes background noise while capturing low and high-frequency information in voice signals, subsequently normalising the audio signals. Difficulties in generalisation and lack of interpretability pose obstacles to extracting relevant characteristics from speech. To tackle these challenges, an Auto Cepstral Fusion extraction technique was proposed to extract pertinent features, aiming to capture both temporal and spectral characteristics caused by background voice. Moreover, feature selection is crucial to ensure robust classification. To address this, the Principal Stochastic Embedding technique handles high-dimensional data, reduces the influence of noise, and minimises dimensionality. Utilising the XGBoost classifier, the proposed method distinguishes between depressed and non-depressed individuals using the DAIC-WOZ Datasets from USC. The proposed approach achieved a remarkable accuracy of 97.02% and a low MAE of 0.9, positioning it as a promising tool for autonomous depression detection. This proposed model provided enhanced accuracy by effectively integrating multiple data modalities. Developing advanced machine learning methods presents interesting chances to improve depression detection systems. In particular, deep learning has demonstrated great promise for identifying complicated patterns and characteristics in large, complex data sets. Future studies might examine how deep learning

architectures can be applied to assess multimodal data more effectively and accurately identify minor signs of depression. By utilising wearable technology, smartphone apps, and other digital platforms, these systems might continually monitor people's physiological and behavioral signals and offer therapies and notifications when early indications of depression are identified. These kinds of systems can lessen the harmful effects of depression and enhance people's general well-being by providing early detection and access to mental health services.

References

1. Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). World Health Organization. 2017. 22 p.
2. Uddin M.Z., Dysthe K.K., Følstad A., Brandtzaeg P.B. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications. 2022. vol. 34(1). pp. 721–744.
3. Jacobson N.C., Chung Y.J. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors. 2020. vol. 20(12). DOI: 10.3390/s20123572.
4. Ormel J., Kessler R.C., Schoevers R. Depression: More treatment but no drop in prevalence: how effective is treatment? And can we do better? Current opinion in psychiatry. 2019. vol. 32(4). pp. 348–354.
5. Culpepper L. Understanding the burden of depression. The Journal of Clinical Psychiatry. 2011. vol. 72(6). DOI: 10.4088/JCP.10126tx1c.
6. Sadock B.J., Sadock V.A., Ruiz P. Compêndio de Psiquiatria: Ciência do Comportamento e Psiquiatria Clínica. Artmed Editora. 2016. 1490 p.
7. Mundt J.C., Vogel A.P., Feltner D.E., Lenderking W.R. Vocal acoustic biomarkers of depression severity and treatment response. Biological psychiatry. 2012. vol. 72(7). pp. 580–587.
8. Hashim N.W., Wilkes M., Salomon R., Meggs J., France D.J. Evaluation of voice acoustics as predictors of clinical depression scores. Journal of Voice. 2017. vol. 31(2). DOI: 10.1016/j.jvoice.2016.06.006.
9. Khoo L.S., Lim M.K., Chong, C.Y., McNaney R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020348.
10. Low D.M., Bentley K.H., Ghosh S.S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology. 2020. vol. 5(1). pp. 96–116.
11. Asci F., Costantini G., Di Leo P., Zampogna A., Ruoppolo G., Berardelli A., Saggio G., Suppa A. Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors. 2020. vol. 20(18). DOI: org/10.3390/s20185022.
12. Chen Z.S., Galatzer-Levy I.R., Bigio B., Nasca C., Zhang Y. Modern views of machine learning for precision psychiatry. Patterns. 2022. vol. 3(11).
13. Jiang H., Hu B., Liu Z., Wang G., Zhang L., Li X., Kang H. Detecting depression using an ensemble logistic regression model based on multiple speech features. Computational and mathematical methods in medicine. 2018. vol. 1. DOI: 10.1155/2018/6508319.

14. Na K.S., Cho S.E., Geem Z.W., Kim Y.K. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. Neuroscience Letters. 2020. vol. 721. DOI: 10.1016/j.neulet.2020.134804.
15. Hochman E., Feldman B., Weizman A., Krivoy A., Gur S., Barzilay E., Gabay H., Levy J., Levinkron O., Lawrence G. Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. Depression and anxiety. 2021. vol. 38(4). pp. 400–411.
16. Narziev N., Goh H., Toshnazarov K., Lee S.A., Chung K.M., Noh Y. STDD: Short-term depression detection with passive sensing. Sensors. 2020. vol. 20(5). DOI: 10.3390/s20051396.
17. Ware S., Yue C., Morillo R., Lu J., Shang C., Kamath J., Bamis A., Bi J., Russell A., Wang B. Large-scale automatic depression screening using meta-data from wifi infrastructure. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018. vol. 2(4). pp. 1–27.
18. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning. medRxiv. 2020. DOI: 10.1101/2020.06.23.20138651.
19. Qureshi S.A., Hasanuzzaman M., Saha S., Dias G. The Verbal and Non Verbal Signals of Depression--Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019. arXiv preprint arXiv:1904.07656.
20. Chen X., Pan Z. A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. International Journal of Environmental Research and Public Health. 2021. vol. 18(12). DOI: 10.3390/ijerph18126441.
21. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning – an exploratory study. Research on Biomedical Engineering. 2021. vol. 37. pp. 53–64.
22. Baek J.W., Chung K. Context deep neural network model for predicting depression risk using multiple regression. IEEE Access. 2020. vol. 8. pp. 18171–18181.
23. Zogan H., Razzak I., Wang X., Jameel S., Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022. vol. 25(1). pp. 281–304.
24. Effati-Daryani F., Zarei S., Mohammadi A., Hemmati E., Ghasemi Yngyknd S., Mirghafourvand M. Depression, stress, anxiety and their predictors in Iranian pregnant women during the outbreak of COVID-19. BMC psychology. 2020. vol. 8. pp. 1–10.
25. Gratch J., Artstein R., Lucas G.M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P. The distress analysis interview corpus of human and computer interviews. LREC. 2014. pp. 3123–3128.
26. Yang L., Jiang D., Xia X., Pei E., Oveneke M.C., Sahli H. Multimodal measurement of depression using deep learning models. Proceedings of the 7th annual workshop on audio/visual emotion challenge. 2017. pp. 53–59.
27. Yang L., Jiang D., Sahli H. Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access. 2020. vol. 8. pp. 24033–24045.
28. Lu J., Liu B., Lian Z., Cai C., Tao J., Zhao Z. Prediction of Depression Severity Based on Transformer Encoder and CNN Model. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE. 2022. pp. 339–343.
29. Fang M., Peng S., Liang Y., Hung C.C., Liu S. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control. 2023. vol. 82. DOI: 10.1016/j.bspc.2022.104561.

30. Ishimaru M., Okada Y., Uchiyama R., Horiguchi R., Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. Diagnostics. 2023. vol. 13(4). DOI: 10.3390/diagnostics13040727.
31. Cao Y., Hao Y., Li B., Xue J. Depression prediction based on BiAttention-GRU. Journal of Ambient Intelligence and Humanized Computing. 2022. vol. 13(11). pp. 5269–5277.
32. Yin F., Du J., Xu X., Zhao L. Depression detection in speech using transformer and parallel convolutional neural networks. Electronics. 2023. vol. 12(2). DOI: 10.3390/electronics12020328.
33. Ahmed S., Yousuf M.A., Monowar M.M., Hamid M.A., Alassafi M. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2023.3315243.

Jacob Jithin — Research scholar, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education. Research interests: computer science, engineering, machine learning, voice analysis, cyber security. The number of publications — 7. jithu7771@gmail.com; Srivilliputhur, Tamil Nadu, 626126, Krishnankoil, India; office phone: +91(0474)257-7958.

Kannan K.S. — Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education. Research interests: machine learning, deep learning. The number of publications — 20. k.s.kannan@klu.ac.in; Srivilliputhur, Tamil Nadu, 626126, Krishnankoil, India; office phone: +91(04563)289-042.

УДК 004 **DOI** 10.15622/ia.23.6.7

Д. Джейкоб, К. Каннан

УСОВЕРШЕНСТВОВАННАЯ СИСТЕМА МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОНОМНОГО ОБНАРУЖЕНИЯ ДЕПРЕССИИ С ИСПОЛЬЗОВАНИЕМ МОДУЛИРОВАННОГО ВЕЙВЛЕТ-КЕПСТРАЛЬНОГО СЛИЯНИЯ И СТОХАСТИЧЕСКОГО ВСТРАИВАНИЯ

Джейкоб Д., Каннан К. **Усовершенствованная система машинного обучения для автономного обнаружения депрессии с использованием модулированного вейвлет-кепстрального слияния и стохастического встраивания.**

Аннотация. Депрессия – это распространенное психическое заболевание, требующее систем автоматического обнаружения из-за своей сложности. Существующие методы машинного обучения сталкиваются с проблемами, такими как чувствительность к фоновому шуму, медленная скорость адаптации и несбалансированные данные. Для устранения этих ограничений в этом исследовании предлагается новая структура модулированного вейвлет-кепстрального слияния и стохастическая структура встраивания для прогнозирования депрессии. Затем техника модулированных волновых функций удаляет фоновый шум и нормализует аудиосигналы. Трудности с обобщением, которые приводят к отсутствию интерпретируемости, затрудняют извлечение соответствующих характеристик речи. Для решения этих проблем используется автоматическое кепстральное слияние, которое извлекает соответствующие характеристики речи, захватывая временные и спектральные характеристики, вызванные фоновым голосом. Выбор признаков становится важным, когда выбираются релевантные признаки для классификации. Выбор нерелевантных признаков может привести к переобучению, нарушению размерности и меньшей устойчивости к шуму. Поэтому метод стохастической иммерсии справляется с высокоразмерными данными, минимизируя влияние шума и размерности. Кроме того, классификатор XGBoost отличает людей с депрессией и людей без депрессии. В результате предложенный метод использует набор данных DAIC-WOZ Университета Южной Калифорнии для обнаружения депрессий, достигая точности 97,02%, прецизионности 97,02%, полноты 97,02%, оценки F1 97,02%, среднеквадратической ошибки 2,00 и средней абсолютной ошибки 0,9, делая его многообещающим инструментом для автономного обнаружения депрессии.

Ключевые слова: обнаружение депрессии, машинное обучение, ModWave Cepstral Fusion, фоновый шум, классификатор XGBoost, набор данных DAIC-WOZ, автономная система обнаружения, точность.

Литература

1. Depression and other common mental disorders: global health estimates (No. WHO/MSD/MER/2017.2). World Health Organization. 2017. 22 p.
2. Uddin M.Z., Dysthe K.K., Følstad A., Brandtzaeg P.B. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications. 2022. vol. 34(1). pp. 721–744.
3. Jacobson N.C., Chung Y.J. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors. 2020. vol. 20(12). DOI: 10.3390/s20123572.

4. Ormel J., Kessler R.C., Schoevers R. Depression: More treatment but no drop in prevalence: how effective is treatment? And can we do better? Current opinion in psychiatry. 2019. vol. 32(4). pp. 348–354.
5. Culpepper L. Understanding the burden of depression. The Journal of Clinical Psychiatry. 2011. vol. 72(6). DOI: 10.4088/JCP.10126tx1c.
6. Sadock B.J., Sadock V.A., Ruiz P. Compêndio de Psiquiatria: Ciência do Comportamento e Psiquiatria Clínica. Artmed Editora. 2016. 1490 p.
7. Mundt J.C., Vogel A.P., Feltner D.E., Lenderking W.R. Vocal acoustic biomarkers of depression severity and treatment response. Biological psychiatry. 2012. vol. 72(7). pp. 580–587.
8. Hashim N.W., Wilkes M., Salomon R., Meggs J., France D.J. Evaluation of voice acoustics as predictors of clinical depression scores. Journal of Voice. 2017. vol. 31(2). DOI: 10.1016/j.jvoice.2016.06.006.
9. Khoo L.S., Lim M.K., Chong, C.Y., McNaney R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020348.
10. Low D.M., Bentley K.H., Ghosh S.S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope investigative otolaryngology. 2020. vol. 5(1). pp. 96–116.
11. Asci F., Costantini G., Di Leo P., Zampogna A., Ruoppolo G., Berardelli A., Saggio G., Suppa A. Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors. 2020. vol. 20(18). DOI: org/10.3390/s20185022.
12. Chen Z.S., Galatzer-Levy I.R., Bigio B., Nasca C., Zhang Y. Modern views of machine learning for precision psychiatry. Patterns. 2022. vol. 3(11).
13. Jiang H., Hu B., Liu Z., Wang G., Zhang L., Li X., Kang H. Detecting depression using an ensemble logistic regression model based on multiple speech features. Computational and mathematical methods in medicine. 2018. vol. 1. DOI: 10.1155/2018/6508319.
14. Na K.S., Cho S.E., Geem Z.W., Kim Y.K. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. Neuroscience Letters. 2020. vol. 721. DOI: 10.1016/j.neulet.2020.134804.
15. Hochman E., Feldman B., Weizman A., Krivoy A., Gur S., Barzilay E., Gabay H., Levy J., Levinkron O., Lawrence G. Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study. Depression and anxiety. 2021. vol. 38(4). pp. 400–411.
16. Narziev N., Goh H., Toshnazarov K., Lee S.A., Chung K.M., Noh Y. STDD: Short-term depression detection with passive sensing. Sensors. 2020. vol. 20(5). DOI: 10.3390/s20051396.
17. Ware S., Yue C., Morillo R., Lu J., Shang C., Kamath J., Bamis A., Bi J., Russell A., Wang B. Large-scale automatic depression screening using meta-data from wifi infrastructure. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018. vol. 2(4). pp. 1–27.
18. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning. medRxiv. 2020. DOI: 10.1101/2020.06.23.20138651.
19. Qureshi S.A., Hasanuzzaman M., Saha S., Dias G. The Verbal and Non Verbal Signals of Depression--Combining Acoustics, Text and Visuals for Estimating Depression Level. 2019. arXiv preprint arXiv:1904.07656.
20. Chen X., Pan Z. A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health. International Journal of

Environmental Research and Public Health. 2021. vol. 18(12). DOI: 10.3390/ijerph18126441.
21. Espinola C.W., Gomes J.C., Pereira J.M.S., dos Santos W.P. Detection of major depressive disorder using vocal acoustic analysis and machine learning – an exploratory study. Research on Biomedical Engineering. 2021. vol. 37. pp. 53–64.
22. Baek J.W., Chung K. Context deep neural network model for predicting depression risk using multiple regression. IEEE Access. 2020. vol. 8. pp. 18171–18181.
23. Zogan H., Razzak I., Wang X., Jameel S., Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web. 2022. vol. 25(1). pp. 281–304.
24. Effati-Daryani F., Zarei S., Mohammadi A., Hemmati E., Ghasemi Yngyknd S., Mirghafourvand M. Depression, stress, anxiety and their predictors in Iranian pregnant women during the outbreak of COVID-19. BMC psychology. 2020. vol. 8. pp. 1–10.
25. Gratch J., Artstein R., Lucas G.M., Stratou G., Scherer S., Nazarian A., Wood R., Boberg J., DeVault D., Marsella S., Traum D., Rizzo S., Morency L.-P. The distress analysis interview corpus of human and computer interviews. LREC. 2014. pp. 3123–3128.
26. Yang L., Jiang D., Xia X., Pei E., Oveneke M.C., Sahli H. Multimodal measurement of depression using deep learning models. Proceedings of the 7th annual workshop on audio/visual emotion challenge. 2017. pp. 53–59.
27. Yang L., Jiang D., Sahli H. Feature augmenting networks for improving depression severity estimation from speech signals. IEEE Access. 2020. vol. 8. pp. 24033–24045.
28. Lu J., Liu B., Lian Z., Cai C., Tao J., Zhao Z. Prediction of Depression Severity Based on Transformer Encoder and CNN Model. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE. 2022. pp. 339–343.
29. Fang M., Peng S., Liang Y., Hung C.C., Liu S. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control. 2023. vol. 82. DOI: 10.1016/j.bspc.2022.104561.
30. Ishimaru M., Okada Y., Uchiyama R., Horiguchi R., Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. Diagnostics. 2023. vol. 13(4). DOI: 10.3390/diagnostics13040727.
31. Cao Y., Hao Y., Li B., Xue J. Depression prediction based on BiAttention-GRU. Journal of Ambient Intelligence and Humanized Computing. 2022. vol. 13(11). pp. 5269–5277.
32. Yin F., Du J., Xu X., Zhao L. Depression detection in speech using transformer and parallel convolutional neural networks. Electronics. 2023. vol. 12(2). DOI: 10.3390/electronics12020328.
33. Ahmed S., Yousuf M.A., Monowar M.M., Hamid M.A., Alassafi M. Taking all the factors we need: A multimodal depression classification with uncertainty approximation. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2023.3315243.

Джейкоб Джитин — научный сотрудник, факультет информатики и инженерии, Каласалингамская академия исследований и образования. Область научных интересов: компьютерные науки, инженерия, машинное обучение, анализ голоса, кибербезопасность. Число научных публикаций — 7. jithu7771@gmail.com; Шривиллипутур, Тамил Наду, 626126, Кришнанкойл, Индия; р.т.: +91(0474)257-7958.

Каннан К.С. — профессор, факультет информатики и инженерии, Каласалингамская академия исследований и образования. Область научных интересов: машинное обучение, глубокое обучение. Число научных публикаций — 20. k.s.kannan@klu.ac.in; Шривиллипутур, Тамил Наду, 626126, Кришнанкойл, Индия; р.т.: +91(04563)289-042.

УДК 004.89:004.04 **DOI** 10.15622/ia.23.6.8

Г.В. Дорохина

ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ КАК ЗАДАЧА КЛАССИФИКАЦИИ РЯДОВ НА МНОЖЕСТВЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ЭЛЕМЕНТОВ СЛОЖНЫХ ОБЪЕКТОВ С ПРИМЕНЕНИЕМ УСОВЕРШЕНСТВОВАННОГО TRIE-ДЕРЕВА

Дорохина Г.В. **Пофонемное распознавание как задача классификации рядов на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева.**

Аннотация. Последовательности, в том числе последовательности векторов, применимы в любых предметных областях. Последовательности скалярных значений или векторов (ряды) могут быть порождены последовательностями более высокого порядка, например: последовательностями состояний, элементов сложных объектов. Работа посвящена применению усовершенствованного trie-дерева в задаче классификации ряда на множестве последовательностей элементов сложных объектов методом динамического программирования. Рассмотрены сферы применения динамического программирования. Показано, что динамическое программирование приспособлено к многошаговым операциям вычисления аддитивных (мультипликативных) мер подобия / различия. Утверждается, что усовершенствованное trie-дерево применимо в задаче классификации ряда на множестве последовательностей элементов сложных объектов методом динамического программирования при использовании таких мер подобия / различия. Выполнен анализ иерархических представлений множеств последовательностей. Описаны преимущества, которые обеспечивает усовершенствованное trie-дерево по сравнению с традиционными представлениями других сильноветвящихся деревьев. Разработано формальное описание усовершенствованного trie-дерева. Дано пояснение ранее полученным данным о существенном приросте скорости операций добавления и удаления последовательностей в усовершенствованном trie-дереве относительно использования массива с индексной таблицей (24 и 380 раз, соответственно). Выполнена постановка задачи пофонемного распознавания речевых команд как задачи классификации ряда на множестве последовательностей элементов сложных объектов и изложен метод её решения. Разработан метод классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева. Он исследован на примере пофонемного распознавания с иерархическим представлением словаря классов речевых команд. В этом методе распознавание речевых команд выполняют в процессе обхода усовершенствованного trie-дерева, хранящего множество транскрипций речевых команд – последовательностей транскрипционных символов, которые обозначают классы звуков. Численные исследования показали, что классификация ряда как последовательности элементов сложных объектов повышает частоту правильной классификации по сравнению с классификацией ряда на множестве рядов, а применение усовершенствованного trie-дерева сокращает затраты времени на классификацию.

Ключевые слова: trie-дерево, множество последовательностей, классификация рядов на множестве последовательностей элементов сложных объектов, динамическое программирование, пофонемное распознавание речевых команд.

1. Введение. Модели структурирования данных для скоростного поиска строк разрабатывают с конца 50-х годов XX века. Систематизировали и развивали методы представления данных А. Ахо и Дж. Ульман, Н. Вирт [1], Д. Кнут [2], Р. Бриандe [3], Д. Гасфилд [4].

Структурой данных, обеспечивающей наибольшую скорость поиска во множестве последовательностей, является дерево цифрового поиска – то же, что лучевой поиск Р. Бриандe [3], trie-дерево, «префиксное дерево», «бор». Особенность организации trie-дерева состоит в том, что элементы последовательности размещены вдоль ветвей дерева. Допустим, задано множество последовательностей символов. Пусть у некоторых из этих последовательностей n начальных символов совпадают со строкой s длины n. В trie-дереве начальные символы указанных последовательностей представлены фрагментом ветви из n вершин. При поиске строки s в указанном множестве её сопоставление с общей начальной частью множества последовательностей происходит однократно. То есть, trie-дерево является перспективной структурой данных для обработки множеств последовательностей, о чем свидетельствует возобновление внимания к этой структуре данных.

Последнее время область применения методов анализа последовательностей значительно расширилась. В виде последовательности представляют информацию о состоянии объектов различной природы. Анализ последовательностей используют в социальных науках [5] и здравоохранении [6, 7], в биологии [4], в экономической сфере [8 – 10], в технике [11 – 13] и т.д.

Многие реальные приложения, такие как веб-майнинг, анализ текста, биоинформатика, системная диагностика и распознавание действий, имеют дело с последовательными данными. В этих приложениях решается задача поиска содержательных шаблонов или создания эффективных прогностических моделей [14], для чего применяют методы машинного обучения, как с учителем, так и без, например, K-средних или метод опорных векторов (SVM). Известное решение в области интеллектуального анализа данных заключается в использовании последовательных шаблонов [15]. Этот подход сначала добывает последовательные шаблоны из набора данных, а затем представляет каждую последовательность в наборе данных как вектор признаков с двоичными компонентами, указывающими, содержит ли эта последовательность конкретный последовательный шаблон. Число последовательных шаблонов часто очень велико. Это

приводит к проблемам с высокой размерностью и разреженностью данных [15].

В некоторых случаях необходимо выполнить нечеткий поиск последовательности во множестве – определить наиболее сходную или отличающуюся последовательность по отношению ко входной. Нечеткий поиск в данном случае можно рассматривать как задачу классификации на множестве последовательностей. Примеры нечеткого поиска в словаре строк: проверка орфографии, распознавание текстов, обнаружение фейковых сайтов [16], поиск нежелательного контента в социальных сетях, поиск дублирующихся данных в базах данных, поиск цитирований и др. Сходство между последовательностями определяют попарно [17, 18]. Для пар строк применяют расстояния Хемминга, Левенштейна, Дамерау-Левенштейна. Обзору мер расхождения между последовательностями посвящена работа [19].

Классические методы сопоставления пар последовательностей разной длины основаны на динамическом программировании. С его помощью последовательности «выравнивают» и сопоставляют, в результате чего определяют их «меру подобия» или «меру различия».

Динамическое программирование относят к перечислительным техникам точных методов поиска решений оптимизационных задач. Это «метод оптимизации, приспособленный к операциям, в которых процесс принятия решений может быть разбит на отдельные этапы (шаги)» [20], – многошаговым операциям. Основоположник метода – Р. Беллман. Методом динамического программирования «могут решаться задачи, приводящиеся к сетевым моделям», например: «транспортные задачи с произвольной … функцией затрат; задачи замены оборудования…; задачи управления запасами и другие» [20].

Динамическое программирование как метод многокритериальной оптимизации [21] заменяет одновременный выбор значений большого числа переменных решаемой экстремальной задачи поочередным определением каждой из них. Выбор значений переменных трактуют как многоэтапный процесс управления дискретной системой, имеющей конечное множество состояний, и управляемой в дискретном времени (пошагово) путём применения воздействий из конечного множества, в результате чего изменяется состояние системы. Определены начальное и множество финальных состояний системы, с изменением состояния системы связан платеж. Последовательности управлений, переводящие систему из начального

в одно из финальных состояний, определяют полные траектории движения системы. Задача состоит в поиске полной траектории, оптимальной по значению суммарного платежа [21].

Динамическое программирование используют как для решения статических задач, например, связанных с распределением ресурсов, так и для задач, связанных с динамикой процесса или системы, применяют для решения задач управления [22]. Упрощение процесса решения достигают за счет ограничения области и количества вариантов, исследуемых при переходе к очередному этапу [22].

Недостатками динамического программирования (ДП) считают следующие проблемы [22]. «Проблема ДП-1» состоит в отсутствии единого универсального метода решения, поскольку каждая задача имеет свои особенности и требует поиска приемлемой совокупности методов решения. «Проблема ДП-2» связана с большими объемами и трудоемкостью решения многошаговых задач, имеющих множество состояний, что приводит к необходимости отбора задач малой размерности либо использования сжатой информации.

Длительное время динамическое программирование применяли в задачах анализа текста и речи, где данные представимы в виде последовательностей разной длины. В задачах обработки символьных последовательностей метод динамического программирования развивали: В.И. Левенштейн, Ф. Дамерау, Р. Вагнер и М. Фишер, Д. Кнут, Д. Моррис и В. Пратт; Р. Бойер и Дж. Мур, А. Ахо и М. Корасик, С. Нидлман и К. Вунш, П. Смит, М. Ватерман; в задачах обработки речи – Т.К. Винцюк [23], Х. Сакое и С. Чиба, Л. Рабинер, В.Ю. Шелепов [24]. Многие методы и алгоритмы, которые зарождались и развивались в связи с задачами распознавания речи, сейчас находят применение в других областях – робототехника, медицина, биоинформатика и др.

Сопоставление и выравнивание пары последовательностей разной длины реализует алгоритм нелинейного растяжения-сжатия временной оси (Dynamic Time Warping, DTW). Его в настоящее время применяют [25 – 29] не только в распознавании речи и строковых алгоритмах, но и в вычислительной геометрии, в различных областях для анализа упорядоченных по времени данных, включая видео, аудиофайлы, временные ряды, измерения и отслеживание данных GPS. Алгоритм DTW используют при классификации и кластеризации последовательностей, например, в задачах распознавания жестов и верификации подписи, в беспроводной связи, в интеллектуальном анализе данных и др. Применительно

к распознаванию речи этот метод используют при недостатке обучающих данных.

Работы автора [30 – 32] посвящены распознаванию речи на основе динамического программирования. К задачам распознавания речи относят: идентификацию дикторов, распознавание изолированных (отдельно произносимых) слов, распознавание речевых команд, распознавание дискретной речи (распознавание речи при пословной диктовке) и выделение ключевых слов в потоке слитной речи, распознавание слитной речи.

Проблема распознавания образов является одной из центральных проблем искусственного интеллекта. При этом «понятие образа можно определить так: объекты, для которых выполняется отношение эквивалентности (одинаковые объекты) или, по крайней мере, отношение толерантности (похожие объекты), в своей совокупности составляет образ» [33]. Проблема распознавания образов «возникла при изучении физиологических свойств мозга», а именно его способности «отвечать на бесконечное множество состояний внешней среды конечным числом реакций» [33]. Распознавание включает в себя обучение опознаванию и опознавание образов. Обучение опознаванию включает выделение классов, выбор пространства признаков образов, разработку классификатора. Опознавание образов состоит в применении классификатора к объекту.

Распознавание изолированных слов является частным случаем задачи распознавания речевых команд. Под речевой командой, с одной стороны, подразумевается произнесение слова или слитное произнесение последовательности слов, которые отделены от остальной речи межфразовыми паузами. С другой стороны, задан перечень слов и последовательностей слов, которые могут быть результатом распознавания. При распознавании оцифрованный звук преобразуют в ряд (последовательность векторов признаков), в результате обработки которого определяют произнесённое слово и последовательность слов. Так как задача сводится к выбору одного из известных классов, то имеем дело с задачей классификации.

Пофонемным называется «распознавание, при котором эталоны слов составляются из общей для всех слов совокупности эталонных элементов, а эталонные элементы интерпретируются как фонемы или части фонем, ответственные за элементарные участки речевого сигнала» [23, 30]. В качестве эталонных элементов могут выступать, например, аллофоны [30, 31] или дифоны [34].

В рамках разработки метода пофонемного распознавания отдельно произносимых слов (2003 г.) на основе trie-дерева автором

создана структура данных для хранения множества строк фонетических транскрипций слов словаря [30]. Пофонемное распознавание выполнялось в процессе обхода этого дерева. Структура данных обеспечивала: 1) быстрый поиск строк; 2) быстрое распознавание слова большого словаря (тысяча слов) за счет однократного сопоставления общих начальных частей транскрипции с соответствующим фрагментом речи; 3) хранение транскрипций в древовидной структуре (возможность получить написание заданной транскрипции из дерева). Возможности древовидной структуры 1) и 3) использованы для хранения больших множеств строк (содержат миллионы строк) и скоростного поиска в них при разработке модуля декларативного морфологического анализа слов русского языка (2004 г.) [35]. В результате этих работ появилось усовершенствованное дерево цифрового поиска [36] – то же, что усовершенствованное trie-дерево. Данная работа является обобщением и системным изложением исследований усовершенствованного trie-дерева и его возможных областей его применения [37 – 39].

Допустим, выделены классы элементов объектов. Назовем образом некоторое обобщённое описание класса в пространстве признаков. Назовём элементы объекта однотипными, если их описания принадлежат одному пространству признаков. В работе [37] предложено ввести алфавит обозначений для классов однотипных элементов. Это позволяет информацию о множестве сложных объектов, представимых как последовательности элементов, хранить и обрабатывать как множество строк.

В работе [37] показано, что trie-дерево [2, 3] – «представление словаря строк для поиска», а усовершенствованное trie-дерево [36] – «представление словаря строк для хранения и поиска». Представление словаря строк для поиска является поисковой структурой, и само множество строк необходимо хранить во внешней по отношению к этой структуре данных памяти. Исследование затрат памяти «представления словаря строк для хранения и поиска» по сравнению с затратами на хранение «представления словаря строк для поиска» одновременно с самим множеством строк выполнено в работе [38].

Работа [39] посвящена, в том числе, точному описанию структур данных усовершенствованного trie-дерева и его исследованию. Опубликованные в ней данные численных исследований скорости выполнения основных операций на множестве словоформ русского языка (около 2 млн. строк) показывают, что у усовершенствованного trie-дерева по сравнению использованием массива с индексной таблицей скорость добавления и удаления строк

возрастает в 24 и 380 раз соответственно. В данной работе этим значениям даны пояснения.

В работе [31] разработан метод пофонемного распознавания речевых команд малого словаря. Некоторые неточности приведенных в ней алгоритмов устранены в данной работе.

В настоящей работе выполнена постановка задачи пофонемного распознавания речевых команд как задачи классификации ряда на множестве последовательностей элементов сложных объектов; разработан метод классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева на примере пофонемного распознавания с иерархическим представлением словаря классов речевых команд.

2. Актуальность и анализ состояния проблемы. Алгоритм DTW применяют к последовательностям одномерных значений или векторов (назовём их рядами). При этом многие ряды могут быть проявлениями последовательностей более высокого порядка. Так, фрагменту ряда может соответствовать некоторое состояние объекта наблюдения, а ряду – последовательность состояний (определённый сценарий). Или фрагмент ряда может описывать некоторый элемент объекта (звук слова), а ряд – последовательность элементов – сложный объект (слово). Такое рассмотрение рядов особенно актуально в связи с активным развитием направлений извлечения закономерностей и знаний из последовательностей и из процессов [40], работами по классификации и кластеризации последовательностей [41 – 43], в том числе последовательностей сложных объектов.

Когда на множестве последовательностей сложных объектов выбирают наилучшую в соответствии с заданной целевой функцией последовательность, имеем дело с задачей оптимизации. Когда нужно выбрать класс сложных объектов, которому в наибольшей степени принадлежит рассматриваемый объект, решают задачу классификации. Если мера подобия / различия сложного объекта представима в виде суммы или произведения мер расхождения для элементов объекта, то для её вычисления можно применить динамическое программирование.

Актуально рассмотреть усовершенствованное trie-дерево как способ представления множества последовательностей элементов сложных объектов для решения задач классификации на множестве таких последовательностей при использовании аддитивной или мультипликативной меры подобия / различия. Это снижает вычислительную сложность за счет однократного вычисления меры

подобия / различия для одинаковых начальных фрагментов последовательностей, и тем самым решает «проблему ДП-2». Можно предположить аналогичное применение усовершенствованного trie-дерева для задач оптимизации на множестве последовательностей элементов сложных объектов при использовании аддитивной или мультипликативной целевой функции, что может быть предметом последующих работ.

В связи с разнообразием задач, которые решают методом динамического программирования, трудно проанализировать все аспекты отсутствия единого универсального метода решения задач методом динамического программирования («проблема ДП-1»). При этом актуальным является рассмотрение усовершенствованного trie-дерева как универсального способа представления и множества рядов, и множества последовательностей элементов сложных объектов, который обеспечивает:

– скоростной поиск в большом множестве последовательностей, в том числе нечёткий поиск с применением динамического программирования в задачах классификации;

– быструю модификацию иерархической структуры и минимальные затраты памяти на ее хранение.

3. Цель и задачи работы. Целью работы является характеристика усовершенствованного trie-дерева [36] как способа представления множества последовательностей элементов сложных объектов в задаче классификации ряда на этом множестве методом динамического программирования.

В работе поставлены и решены следующие задачи:

– выполнен анализ иерархических представлений множеств последовательностей;

– сформулированы две новые проблемы сильноветвящихся деревьев, которые решает усовершенствованное trie-дерево;

– сформулированы проблемы сопоставления рядов алгоритмом Dynamic Time Warping;

– разработано формальное описание усовершенствованного trie-дерева с учетом решения проблем сильноветвящихся деревьев;

– выполнена постановка задачи пофонемного распознавания речевых команд как задачи классификации ряда на множестве последовательностей элементов сложных объектов с применением алгоритма Dynamic Time Warping и изложен метод её решения;

– разработан метод классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева на примере пофонемного

распознавания с иерархическим представлением словаря классов речевых команд.

4. Анализ иерархических представлений множеств последовательностей. Иерархические структуры позволяют быстро искать данные, в связи с чем их используют для создания индексов или специальных поисковых структур. Деревья поиска иерархически разбивают пространство поиска на области в соответствии с некоторыми условиями. Например, предикат корневой вершины бинарного дерева поиска разбивает множество ключей на два подмножества: ключи, значение которых меньше хранимого в вершине, и ключи, значение которых, соответственно, больше. Такое разбиение продолжается рекурсивно до тех пор, пока в подмножестве не останется один ключ.

Как представления множеств последовательностей, деревья можно разделить на две группы, исходя из объема данных, которые подвергают анализу в каждой вершине. К первой принадлежат деревья, в которых вся последовательность является ключом и подлежит анализу в каждой вершине. Примерами могут служить бинарное дерево поиска и сильноветвящиеся деревья [1]: (a,b)-дерево, b-дерево, b^+ дерево, R-дерево, k-d-дерево. К этой же группе принадлежат деревья решений, вершина которых содержит определенный предикат. Тогда как, например, в бинарном дереве поиска во всех вершинах применяют одинаковые предикаты «равно ключу в вершине», «меньше (больше) ключа в вершине», в дереве решения каждой вершине назначен собственный предикат.

Ко второй группе принадлежат деревья, в вершинах которых анализируют часть ключа (один или несколько элементов последовательности). При этом элементы последовательности, которые анализируют в некоторой вершине, предшествуют элементам последовательности, которые анализируют в её вершинах-потомках. К этой группе относят trie-дерево, префиксное и суффиксное дерево, дерево из алгоритма Ахо-Корасик.

Деревья второй группы применяют, в основном, для обработки строк и последовательностей. В то же время перспективным представляется их использование и для других классов последовательностей, например последовательностей образов (классов объектов) [37], элементов сложных объектов.

Trie-дерево является инструментом поиска во множестве последовательностей. Хранение множества последовательностей, при необходимости, организуют во внешней по отношению к древовидной структуре дополнительно выделенной области памяти либо в листовых

вершинах. Тогда как возможность хранения последовательностей внутри древовидной структуры позволила бы создавать словари, обеспечивающие: скоростной поиск; повышение скорости добавления и удаления элементов; сокращение затрат памяти на хранение.

Рассмотрим trie-дерево [2, 3]. В этой структуре максимальное число потомков вершины ограничено размером алфавита, которому принадлежат строки, а отдельной строке соответствует ветвь от корня к листовой вершине. Trie-дерево имеет:

– внутренние вершины, у которых как минимум два потомка;
– листья, в которых хранятся последовательности [44].

Trie-дерево является сильноветвящимся деревом.

5. Проблемы сильноветвящихся деревьев. Одной из проблем сильноветвящихся деревьев является проблема размера вершины. Допустим, дерево цифрового поиска хранит последовательности, элементы которых принадлежат алфавиту A из τ элементов. Если в каждой вершине для ссылки на дочерние элементы хранить массив из τ элементов (по одному для каждого элемента алфавита), то «лишние» затраты памяти будут чрезмерно большими, так как большая часть будет «пустыми указателями». Если же в массиве ссылок на дочерние элементы хранить только «непустые» ссылки, то вершины дерева будут отличаться друг от друга по размеру. А это затрудняет хранение вершин в структурах данных с произвольным доступом (например, при их хранении в файле).

Ещё одной проблемой описания древовидных структур в классических и современных учебниках является использование для ссылки на вершины аппарата указателей (ссылочных структур). Например, узел дерева бинарного поиска [2] описан структурой NodeDesc, приведенной в листинге 1.

```
TYPE Node = POINTER TO NodeDesc;
TYPE NodeDesc = RECORD op: CHAR; left, right: Node END;
```

Листинг 1. Структура, описывающая дерево бинарного поиска

С деревьями на основе аппарата указателей проблематично выполнять операции без их загрузки в оперативную память компьютера. В то же время, в теории графов общепринятая практика – перенумеровать вершины и ссылаться на них по номеру. В графовых базах данных для ссылки на вершину также используют идентификаторы вершин.

6. Проблемы сопоставления рядов алгоритмом Dynamic Time Warping на примере задачи распознавания речевых команд.

В рамках этой задачи фактически решают задачу классификации ряда на основе меры расхождения распознаваемого и эталонного рядов. Алгоритм сопоставляет распознаваемую команду со всеми эталонами речевых команд. При этом распознаваемая речевая команда R и каждый эталон E являются последовательностями векторов признаков (рядами):

$$E = \mathbf{e}_1, ..., \mathbf{e}_j, ..., \mathbf{e}_n, \; R = \mathbf{r}_1, ..., \mathbf{r}_i, ..., \mathbf{r}_m, \tag{1}$$

полученными в результате предварительной обработки оцифрованного звука, который соответствует произнесению речевой команды.

Меру расхождения $d(E,R)$ находят следующим образом [24]. Вычисляют матрицу расстояний $\mathbf{D}$ между элементами $\mathbf{e}_i$ и $\mathbf{r}_j$:

$$\mathbf{D} = \left\| d_{ij} \right\|_{n \times m}, \; d_{ij} = d(e_i, r_j). \tag{2}$$

Далее, по матрице $\mathbf{D}$ вычисляют элементы DTW-матрицы выравнивания $\mathbf{K} = \left\| k_{ij} \right\|_{n \times m}$ по рекуррентным формулам:

$$\begin{aligned} & k_{11} = d_{11}; \\ & k_{1j} = d_{1j} + k_{1(j-1)}, j = \overline{2,n}; \\ & k_{i1} = d_{i1} + k_{(i-1)1}, i = \overline{2,m}; \\ & k_{ij} = d_{ij} + \min(k_{(i-1)(j-1)}, k_{i(j-1)}, k_{(i-1)j}), i = \overline{2,m}, \; j = \overline{2,n}. \end{aligned} \tag{3}$$

Мерой расхождения между E и R является значение k_{nm}:

$$d(E,R) = k_{nm}. \tag{4}$$

Полученная DTW-матрица $\mathbf{K}$, позволяет определить путь выравнивания – множество пар индексов соответствующих друг другу элементов $\mathbf{e}_{p_h}$ и $\mathbf{r}_{q_h}$ рядов E и R [24]:

$$M = \{(p_h, q_h)\}_{h=1}^{H}, \tag{5}$$

где H – количество пар в пути выравнивания. Эти пары определяют итеративно по формулам [24]:

$$\begin{aligned} &p_H = n, q_H = m \\ &(p_{h-1}, q_{h-1}) = \arg\min_{(p',q')\in P_h}(k_{p'q'}) \\ &P_h = \{(p',q') : p_h - 1 \le p' \le p_h, q_h - 1 \le q' \le q_h, (p' \ne p_h \vee q' \ne q_h)\} \\ &p_1 = 1, q_1 = 1 \end{aligned} . \quad (6)$$

Мера расхождения $d(E,R)$ представима также в виде суммы:

$$d(E,R) = \sum_{h=1}^{H} d(e_{p_h}, r_{q_h}) . \quad (7)$$

Словарь классов речевых команд W – это множество эталонов:

$$W = \left\{E_y\right\}_{y=1}^{z}, \quad (8)$$

а целевая функция находит номер класса y, между эталоном которого E_y и распознаваемой речевой командой R мера расхождения $d(E_y,R)$ минимальна:

$$res = \arg\min_{y=\overline{1,z}} d(E_y,R) . \quad (9)$$

Проблемы, возникающие при сопоставлении рядов с использованием алгоритма DTW [31]:

1) проблема пропуска алгоритмом DTW отличающихся элементов и учёта сходных (её пытаются решать введением ограничений на путь выравнивания, но это не всегда приносит ожидаемый эффект);

2) проблема длин эталонов при использовании DTW: если длина одного из эталонов значительно меньше длин остальных, мера расхождения от него до распознаваемого ряда может оказаться минимальной;

3) проблема вычислительной сложности DTW: вычислительная сложность классификации на основе алгоритма DTW

ряда длины m пропорциональна величине: $m \cdot \sum n_y$, где n_y – длина эталона с номером y в словаре классов речевых команд (8).

Решение проблемы длин эталонов за счёт нормировки меры расхождения по длине диагонали DTW-матрицы предложено в работе [31]. В данной работе предложены решения проблем пропуска алгоритмом DTW отличающихся элементов и учёта сходных и вычислительной сложности DTW.

7. Формальное описание усовершенствованного trie-дерева с учетом решения проблем сильноветвящихся деревьев. Отличия [39] усовершенствованного trie-дерева [36] от trie-дерева [2, 3]:

1) усовершенствованное trie-дерево содержит массив ссылок на вершины окончания строк;

2) вершины усовершенствованного trie-дерева имеют одинаковую структуру, в отличие от разделения на внутренние и листовые вершины у trie-дерева;

3) вершины усовершенствованного trie-дерева содержат ссылки на родительскую вершину и идентификатор строки, оканчивающейся в вершине;

4) в вершинах усовершенствованного trie-дерева хранится ссылка на множество вершин-потомков, что обеспечивает фиксированный размер вершины и позволяет хранить вершины в структурах данных произвольного доступа, а также ссылаться на вершины и множества вершин с помощью их номеров вместо использования аппарата указателей.

Пусть G – множество последовательностей, элементы которых принадлежат алфавиту A символов (или идентификаторов), $|A| = \tau$. Представление G в виде усовершенствованного trie-дерева:

$$\begin{aligned} &(O, Ch, X) \\ &O = \{\mathbf{o}_j\}, \mathbf{o}_j = (o_1^j, o_2^j, o_3^j, o_4^j) \\ &Ch = \{\mathbf{ch}_k\} \\ &X = \{x_i\}, x_i = j : o_2^j = i, i = \overline{1, z}, z = |G| \end{aligned}, \quad (10)$$

где O – массив векторов, описывающих вершины усовершенствованного trie-дерева; $o_1^j \in A$ – элемент последовательности; o_2^j – идентификатор последовательности, завершающейся в вершине номер j; o_3^j – номер в упорядоченном

множестве Ch массива, хранящего номера вершин-потомков вершины $\mathbf{o}_j$; o_4^j – номер родительской вершины для вершины $\mathbf{o}_j$; Ch – упорядоченное множество массивов $\mathbf{ch}_k$; $\mathbf{ch}_k$ – массив номеров вершин-потомков некоторой вершины ($ch_{o_3^j}$ – массив номеров вершин-потомков вершины $\mathbf{o}_j$); X – массив номеров вершин, в которых оканчиваются последовательности; x_i – номер j вершины $\mathbf{o}_j$, у которой поле o_2^j (идентификатор последовательности, завершающейся в вершине номер j) равно i .

Усовершенствованное trie-дерево является представлением словаря строк для хранения и поиска [36], так как для него определены функции получения: 1) идентификатора строки по её написанию; 2) строки по её идентификатору. Структура данных в программной реализации усовершенствованного trie-дерева описана в работе [39].

Фиксированный размер вершины позволяет хранить и обрабатывать вершины дерева в структурах данных с произвольным доступом (например, файлах); обрабатывать деревья, частично загруженные в оперативную память.

В проведенном эксперименте фиксированный размер вершины и использование идентификаторов вершин вместо аппарата указателей на словаре словоформ русского языка, содержащем около 2 млн. уникальных строк, увеличило на 40% скорость загрузки древовидной структуры из файла. Данный результат получен следующим образом. Измерено время загрузки усовершенствованного trie-дерева. Затем усовершенствованное trie-дерево модифицировано – изменён способ ссылки на вершины: вместо использования массивов и идентификаторов вершин применён аппарат указателей. После этого измерено время загрузки модифицированного дерева.

Исследования [38, 39], выполненные на множестве словоформ русского языка (около 2 млн. строк), показали, что предложенная иерархическая структура по сравнению с использованием массива с индексной таблицей уменьшает затраты памяти на 20%, времени на выполнение операций: поиска – на 37%, добавления и удаления – в 24 и 380 раз, соответственно.

Требует пояснения столь существенный прирост скорости операций добавления и удаления. Он связан с тем, что строка, попавшая в trie-дерево, никогда не меняет свой идентификатор. Точно так же, как вершина и массив вершин-потомков. Среди идентификаторов строк, вершин и массивов вершин-потомков есть

предопределённое значение 0, которое указывает на отсутствие ссылки. Значение поля $o_1^j = 0$ указывает на отсутствие символа в вершине $\mathbf{o}_j$. Например, в представлении (10) у корневой вершины $\mathbf{o}_1$ номер родительской вершины $o_4^1 = 0$ и номер символа вершины $o_1^1 = 0$; если в вершине $\mathbf{o}_j$ не оканчивается ни одна строка, то идентификатор строки, завершающейся в вершине номер j, – поле $o_2^j = 0$; если вершина $\mathbf{o}_j$ не имеет потомков, то идентификатор массива вершин-потомков $o_3^j = 0$.

Когда нужно удалить строку с номером i, удаляют фрагмент ветви, которая оканчивается в вершине с номером x_i, затем присваивают $x_i = 0$ и добавляют i в массив идентификаторов удаленных строк *DeletedStrings* [39]. Позже, когда понадобится добавить строку, идентификатор новой строки будет выбран из массива *DeletedStrings*, если он не пуст. Само нулевое значение остаётся в массиве X.

Аналогично, когда нужно удалить массив потомков массивов $\mathbf{ch}_k$, этот массив очищают, ссылку на него в вершине обнуляют и заносят номер k в массив номеров удалённых массивов вершин-потомков *DeletedChArr* [39]. Сам пустой массив $\mathbf{ch}_k$ остаётся во множестве Ch.

Когда нужно удалить вершину с номером j, то вектору $\mathbf{o}_j$ присваивают значение $\mathbf{o}_j = (0,0,0,0)$, предварительно удалив ссылку на неё из массива номеров вершин-потомков, самих вершин-потомков и массива номеров вершин, в которых оканчиваются последовательности. Заносят номер j в массив номеров удалённых вершин *DeletedKnots* [39]. Сам обнулённый вектор $\mathbf{o}_j$ остаётся во множестве O.

То есть, при операциях добавления и удаления строк в усовершенствованном trie-дереве минимизируются операции выделения памяти и перемещения в памяти больших фрагментов данных, как происходит при удалении элемента в начале массива.

8. Пофонемное распознавание речевых команд как задача классификации ряда на множестве последовательностей элементов сложных объектов с применением DTW. Фонетика рассматривает речь, в т. ч. слова и последовательности слов, как

последовательность фонетических единиц. В качестве фонетических единиц в данной работе используем разновидности звучания фонем, которые называют аллофонами и обозначают транскрипционными символами:

$$A = \{a_i\}_{i=1}^{\tau}. \quad (11)$$

Каждому аллофону соответствует класс звуков. Следует отметить, что эти классы звуков достаточно размыты в пространстве акустических признаков и имеют значительные пересечения.

Фонетическая транскрипция – это последовательность $T = t_1,..,t_\kappa,...,t_l$ транскрипционных символов $t_\kappa \in A$, где l – длина транскрипции T. Она обозначает последовательность аллофонов, из которых состоит слово или последовательность слов. Фонетическую транскрипцию речевой команды (один и более вариантов) можно построить по её написанию. То есть, фонетическая транскрипция речевой команды обозначает последовательность классов звуков, которой соответствует произнесение этой команды. При распознавании оцифрованный звук с произнесением речевой команды преобразуют в последовательность векторов признаков (ряд). По этому ряду необходимо найти последовательность классов звуков, составляющих соответствующую речевую команду.

Определение границ элементов в распознаваемом ряду по их границам в эталоне. Проблемой пофонемного распознавания является то, что границы аллофонов речевой команды внутри распознаваемого ряда $R = \mathbf{r}_1,...,\mathbf{r}_i,...,\mathbf{r}_m$ неизвестны. Для определения этих границ (сегментации распознаваемого ряда) предложено выполнять «условную сегментацию» [30] следующим образом. Распознаваемый ряд R сегментируют, исходя из предположения, что он соответствует последовательности классов звуков, которая обозначена транскрипцией T. То есть, каждой транскрипции будет соответствовать свой набор границ сегментов. Границы классов звуков (аллофонов) в R находят последовательно, исходя из предположений: 1) начало R совпадает с началом первого аллофона речевой команды; 2) при известной левой границе κ-го аллофона в R, его правую границу определяют, используя эталоны κ-го и $(\kappa+1)$-го аллофонов и считая, что элемент $\mathbf{r}_j$ ряда R относится к κ-му аллофону до тех пор, пока он не станет «ближе» к эталону $(\kappa+1)$-го аллофона, чем к эталону κ-го.

В работе [31] с помощью пути выравнивания M (4) пары рядов E и R (1), соответствующих одной и той же последовательности аллофонов (в работе [24] – разным произнесениям одного слова), по известному множеству границ начала аллофонов $B = \{b_1,...,b_\kappa,...,b_l\}$ в E получают соответствующие границы $B' = \{b'_1,...,b'_\kappa,...,b'_l\}$ в R.

Традиционно, путь выравнивания M строят при обратном проходе DTW-матрицы (6). Работа [31] предлагает алгоритмы, позволяющие вычислить путь выравнивания и меру расхождения при прямом проходе DTW-матрицы.

Мера расхождения между парой последовательностей векторов признаков, соответствующих одной последовательности классов звуков. Назовём фрагментом $E[x;y]$ ряда E последовательность $\mathbf{e}_x, \mathbf{e}_{x+1}, \ldots, \mathbf{e}_y$. Аналогично определим фрагмент $R[x';y']$. Пусть E и R соответствуют произнесениям слова с транскрипцией $T = t_1,..,t_\kappa,...,t_l$, а фрагменты $E[b_\kappa; b_{\kappa+1}-1]$ и $R[b'_\kappa; b'_{\kappa+1}-1]$ представляют в E и R аллофон t_κ. Это позволяет найти меру расхождения между E и R путём сопоставления их фрагментов, соответствующих отдельным аллофонам. Действительно, величина (4) представима как сумма мер расхождения между фрагментами E и R, соответствующими одному аллофону [31]:

$$d(E,R) = \sum_{k=1}^{l-1} d(E[b_\kappa; b_{\kappa+1}-1], R[b'_\kappa; b'_{\kappa+1}-1]) + d(E[b_l; n], R[b'_l; m]). \quad (12)$$

Мера расхождения с эталонной последовательностью, синтезированной по последовательности классов звуков. Напомним, что по правилу «условной сегментации» для аллофона, не являющегося последним аллофоном речевой команды, его правую границу необходимо определять, используя вектора признаков, относящиеся к κ-му и $(\kappa+1)$-му аллофонам. Т.е., нужно создавать эталоны для пар соседних аллофонов.

Пусть κ-й и $(\kappa+1)$-й аллофоны слова принимают значения g и h из алфавита A, а эталон аллофона g, за которым следует аллофон h, Et_g^h является последовательностью векторов (рядом) из c_{gh} элементов:

$$Et_g^h = \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{c_{gh}-c'}, \ldots, \mathbf{e}_{c_{gh}}. \quad (13)$$

Причём $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{c_{gh}-c'}$ описывает аллофон g и межфонемный переход между аллофонами g и h (эта часть учитывается при подсчёте меры расхождения), а $\mathbf{e}_{c-c'+1}, \ldots, \mathbf{e}_c$ относится к аллофону h. Понадобятся также эталоны последних аллофонов речевых команд. Обозначим эталон аллофона g, являющегося последним в речевой команде, Et_g^{λ}. Он также представлен рядом (13).

Обозначим через $\Diamond$ операцию конкатенации (соединения) фрагментов последовательностей и рядов. Введённые выше обозначения, позволяют рассматривать эталонный ряд E как эталон, синтезированный путём конкатенации фрагментов эталонов не последних аллофонов $Et_{t_\kappa}^{t_{\kappa+1}}[1; c_{t_\kappa t_{\kappa+1}} - c']$, $\kappa = \overline{1, l-1}$ речевой команды $T = t_1, .., t_\kappa, ..., t_l$ с добавлением эталона последнего аллофона Et_g^{λ}:

$$E = Et_{t_1}^{t_2}[1; c_{t_1 t_2} - c'] \Diamond Et_{t_2}^{t_3}[1; c_{t_1 t_2} - c'] \Diamond \ldots \Diamond Et_{t_{l-1}}^{t_l}[1; c_{t_{(l-1)} t_l} - c'] \Diamond E_{t_l}^{\lambda}. \quad (14)$$

Приведём алгоритм вычисления меры расхождения $d_1(Et_{t_\kappa}^{t_{\kappa+1}}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa)$ от эталона $Et_{t_\kappa}^{t_{\kappa+1}}$ (13) до соответствующего фрагмента $R[b'_\kappa; b'_{\kappa+1}]$, с сохранением границы $b'_{\kappa+1}$ начала $(\kappa+1)$-го аллофона и длины u_κ эталона $Et_{t_\kappa}^{t_{\kappa+1}}$ (Алгоритм 1), который является производным от алгоритма, работы [31]. В рамках данного алгоритма длину $c_{t_\kappa t_{\kappa+1}}$ эталона $Et_{t_\kappa}^{t_{\kappa+1}}$ обозначим через c.

Считаем b'_κ – левую границу κ-го аллофона в R, известной. Левую границу $b'_{\kappa+1}$ следующего за ним $(\kappa+1)$-го аллофона в R получим в процессе вычисления искомой меры расхождения. Для вычисления понадобится DTW-матрица $\mathbf{K}$ размерности $c \times m$.

Заполняем её от элемента с индексом $(1,1)$, который является первым «рабочим элементом». Заполнение происходит, пока рабочим элементом не станет элемент $k_{c\,q}$ строки с номером c. При анализе рабочего элемента $k_{p_h\,q_h}$ используем элементы столбцов q_h и (q_h+1).

1. *Инициализация* // вычислить 2 столбца DTW-матрицы $\mathbf{K}$

1.1 $k_{11} := d(\mathbf{e}_1, \mathbf{r}_{b'_\kappa})$, $k_{s1} := d(\mathbf{e}_s, \mathbf{r}_{b'_\kappa}) + k_{(s-1)1}$, $s = \overline{2,c}$

1.2 $k_{12} := d(\mathbf{e}_1, \mathbf{r}_{b'_\kappa+1}) + k_{11}$

1.3 $k_{s2} := d(\mathbf{e}_s, \mathbf{r}_{b'_\kappa+1}) + \min(k_{(s-1)2}, k_{(s-1)1}, k_{s1}), s = \overline{2,c}$

1.4 $f := 1, j := b'_\kappa$, $p := 1, q := b'_\kappa$

2. *Пока* $(p \le c) \wedge (q \le m)$ *выполнять*

2.1 *если* $q = j + 1$ *то* //вычислить $(q+1)$-й столбец матрицы $\mathbf{K}$

2.1.1 $k_{1(q+1)} := d(\mathbf{e}_1, \mathbf{r}_{q+1}) + k_{1q}$

2.1.2 $k_{s(q+1)} := d(\mathbf{e}_s, \mathbf{r}_{q+1}) + \min(k_{(s-1)q}, k_{(s-1)(q+1)}, k_{s(q+1)}), s = \overline{2,c}$

2.2 $f := p, j := q$

2.3 $(p,q) = \arg\min_{(p',q') \in P_{fj}} (k_{p'q'})$, где $P_{fj} = \{(p',q')\}$,

$f \le p' \le f+1, j \le q' \le j+1, p' \le n, q' \le m, (p' = f \vee q' = j)$

2.4 *если* $(p = c) \wedge (q = j)$ *то* $p := f+1$, $q := j+1$

3. *Если* $(p < c) \wedge (q = m)$ // достигли конца последовательности R

3.1 *то* ***Результат:***

$b'_{\kappa+1} := m$ – номер вектора начала $(\kappa+1)$-го аллофона

$u_\kappa := (c - c')$ – длина эталона аллофона t_κ

$d_1(Et^{t_{\kappa+1}}_{t_\kappa}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) := -1$ («отказ от распознавания»)

3.2 *иначе* // достигли конца эталона $Et^{t_{\kappa+1}}_{t_\kappa}$

3.2.1 $f := p, j := q$

3.2.2 *Пока* $(p > 1) \wedge (q > b_\kappa) \wedge (p \ge c - c')$ *выполнять*

$(p,q) = \arg\min_{(p',q') \in P_{fj}} (k_{p'q'})$, $P_{fj} = \{(p',q')\}$, где

$f - 1 \le p' \le f, j - 1 \le q' \le j, p' \ge 1, q' \ge b_\kappa, (p' = f - 1) \vee (q' = j - 1)$

3.2.3 ***Результат:***

$b'_{\kappa+1} := q+1$ – номер вектора начала $(\kappa+1)$-го аллофона

$u_\kappa := (c - c')$ – длина эталона аллофона t_κ

$d_1(Et^{t_{\kappa+1}}_{t_\kappa}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) := k_{pq}$ – искомая мера расхождения

Алгоритм 1. Мера расхождения $d_1(Et^{t_{\kappa+1}}_{t_\kappa}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa)$

Поэтому алгоритм начнём с вычисления первых двух столбцов, далее при каждом изменении номера столбца рабочего элемента будем вычислять значение следующего столбца.

Если на некотором шаге итерации индексы рабочего элемента приняли значение (p,m) и $p < c$, то делаем вывод о несоответствии ряда R и последовательности аллофонов T, поскольку достигли границы R, но не достигли границы не последнего аллофона. В этом случае присвоим искомой мере расхождения значение –1 и будем говорить, что произошёл «отказ от распознавания».

Когда индексы рабочего элемента достигают значение (c,q), то рассмотрен последний вектор эталона пары аллофонов. Вектор эталона $Et_{t_\kappa}^{t_{\kappa+1}}$ с индексом c уже принадлежит аллофону $t_{\kappa+1}$. Значит, вектор $\mathbf{r}_q$ ряда R принадлежит аллофону $t_{\kappa+1}$, т.е. $b'_{\kappa+1} \le q$.

Найдём границу $b'_{\kappa+1}$, для чего выполним обратный проход по DTW-матрице, пока текущим элементом обратного прохода не станет элемент $k_{(c-c')q_1}$. Это и есть искомая мера расхождения, а значение (q_1+1) является – $b'_{\kappa+1}$ левой границей $(\kappa+1)$-го аллофона в R.

В экспериментальной части данной работы [31] для константы c' использовано значение 2.

Алгоритм 1 используют для вычисления меры расхождения от не последнего аллофона речевой команды до фрагмента R.

Опишем меру расхождения от эталона последнего аллофона до фрагмента R. Для этого обозначим паузу, следующую за последним аллофоном слова, специальным символом λ и введём его в алфавит аллофонов A. Обозначим через $d_1'(Et_{t_l}^{\lambda}, R[b_l;m], u_l)$ меру расхождения между эталоном последнего аллофона Et_{tl}^{λ} слова и фрагментом $R[b_l;m]$. Вычислим $d_1'(Et_{t_l}^{\lambda}, R[b_l;m], u_l)$ как значение элемента $k_{(c-c')m}$ DTW-матрицы $\mathbf{K} = \left\| k_{ij} \right\|_{(c-c')\times m}$ по рекуррентным формулам (3); u_l присвоим значение $(c-c')$.

Мера расхождения между распознаваемой последовательностью и последовательностью классов звуков. Выполним постановку задачи пофонемного распознавания как задачи классификации ряда на множестве последовательностей аллофонов и обоснуем возможность её решения.

Как отмечено выше, для определения границы между соседними аллофонами нужны эталоны для пар соседних аллофонов $Et_{t_к}^{t_{к+1}}$ (13), где $t_к$ текущий аллофон, а $t_{к+1}$ – следующий аллофон.

Пусть образ пары аллофонов $\mathrm{E}_{a_i}^{a_j}$ представлен множеством эталонов $Et_{a_i}^{a_j}$, являющихся рядами (13):

$$\mathrm{E}_{a_i}^{a_j} = \left\{ Et_{a_i}^{a_j} \right\},\ Et_{a_i}^{a_j} = \mathbf{e}_1, \ldots, \mathbf{e}_{c-c'}, \ldots \mathbf{e}_c . \tag{15}$$

Отметим, что длина c у разных эталонов может отличаться, значение c' является константой.

Назовём алфавитом образов аллофонов пару упорядоченных множеств $\langle A, \mathrm{E} \rangle$, содержащую множество обозначений аллофонов A с помощью транскрипционных символов и множество образов пар аллофонов (текущий-следующий аллофон) E:

$$\langle A, \mathrm{E} \rangle,\ \ A = \{a_i\}, \mathrm{E} = \{\mathrm{E}_{a_i}^{a_j}\} . \tag{16}$$

Сформируем словарь классов речевых команд W как совокупность алфавита образов аллофонов $\langle A, \mathrm{E} \rangle$ и словаря строк $G = \{T_y\}_{y=1}^{z}$ – множества фонетических транскрипций речевых команд:

$$W = \langle A, \mathrm{E}, G \rangle . \tag{17}$$

При классификации ряда R (1) нужно найти номер y транскрипции T_y, определяющей последовательность образов пар аллофонов, для которой мера расхождения с R минимальна:

$$res = \arg\min_{y=\overline{1,z}} F(T_y, R) : F(T_y, R) \neq -1 , \tag{18}$$

где $F(T_y, R)$ – функционал для вычисления меры расхождения между последовательностью образов пар аллофонов, полученных по фонетической транскрипции T_y, и распознаваемым рядом R.

Функционал $F(T_y, R)$ позволяет выполнить отказ от распознавания – в случае, когда не удалось найти фрагменты для всех образов пар аллофонов из транскрипции T_y в распознаваемом ряду R, он возвращает значение «–1».

Для разработки функционала $F(T, R)$ потребуется ввести функционалы для вычисления меры расхождения между образом пары аллофонов и соответствующим фрагментом ряда R.

Меру расхождения $\Phi(\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}, \mathrm{R}[b'_\kappa; b'_{\kappa+1}], u_\kappa)$ от образа пары аллофонов $\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}$ до соответствующего фрагмента $R[b'_\kappa; b'_{\kappa+1}]$, с сохранением границы $b'_{\kappa+1}$ начала $(\kappa+1)$-го аллофона и длины u_κ того эталона nEt, для которого минимально отношение меры расхождения к длине диагонали DTW-матрицы и не произошло отказа от распознавания, вычисляем следующим образом.

Из эталонов $Et \in \mathrm{E}_{t_\kappa}^{t_{\kappa+1}}$ сформируем множество $\widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}}$ эталонов, для которых при вычислении $d_1(Et, R[b'_\kappa; b'_{\kappa+1}], u_\kappa)$ не происходит отказ от распознавания:

$$\widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}} = \{Et\} : \left(Et \in \mathrm{E}_{t_\kappa}^{t_{\kappa+1}}\right) \wedge \left(d_1(Et, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) \neq -1\right). \tag{19}$$

Из этого множества выберем эталон nEt, с минимальным отношением меры расхождения к длине диагонали DTW-матрицы:

$$nEt = \underset{Et \in \widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}}}{\arg\min} \left(\frac{d_1(Et, R[b'_\kappa; b'_{\kappa+1}], u_\kappa)}{\sqrt{(u_\kappa)^2 + (b'_{\kappa+1} - b'_\kappa)^2}} \right). \tag{20}$$

Если множество $\widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}} = \varnothing$, происходит отказ от распознавания.

$$\Phi(\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) = \begin{cases} d_1(nEt, R[b'_\kappa; b'_{\kappa+1}], u_\kappa), \widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}} \neq \varnothing \\ -1 \qquad\qquad\qquad\qquad , \textit{иначе} \end{cases}, \tag{21}$$

где $\widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}}$ – множество эталонов пар аллофонов (19), для которых при вычислении $d_1(Et, R[b'_\kappa; b'_{\kappa+1}], u_\kappa)$ не происходит отказ от распознавания; nEt – эталон из множества $\widehat{\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}}$, для которого минимально отношение меры расхождения к длине диагонали DTW-матрицы (20).

При вычислении меры расхождения $\Phi'(\mathrm{E}_{t_l}^{\lambda}, \mathrm{R}[\mathrm{b}_1; \mathrm{m}], \mathrm{u}_\kappa)$ от образа пары последнего аллофона и завершающей паузы $\mathrm{E}_{tl}^{\lambda}$ до соответствующего фрагмента $R[b'_l; m]$ отказа от распознавания не происходит. Вычисление меры расхождения с сохранением длины u_l того эталона, для которого минимальна мера расхождения делённая на длину диагонали DTW-матрицы, происходит следующим образом:

$$\Phi'(\mathrm{E}_{tl}^{\lambda}, R[b'_l; m], u_l) = d'_1(Et_{tl}^{\lambda}, R[b'_l; m], u_l),$$
$$Et_{tl}^{\lambda} = \underset{Et_{tl}^{\lambda} \in \mathrm{E}_{tl}^{\lambda}}{\arg\min} \left(\frac{d'_1(Et_{tl}^{\lambda}, R[b'_l; m], u_l)}{\sqrt{(u_l)^2 + (m - b'_\kappa)^2}} \right). \quad (22)$$

Функционал $F(T, R)$ для вычисления меры расхождения между последовательностью образов пар аллофонов, полученных по фонетической транскрипции T и рядом R, позволяющий выполнить отказ от распознавания:

$$F(T,R) = \begin{cases} -1, & \exists\, \kappa = \overline{1, l-1} : \Phi(\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) = -1 \\ \dfrac{\sum_{\kappa=1}^{l-1} \Phi(\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}, R[b'_\kappa; b'_{\kappa+1}], u_\kappa) + \Phi'(\mathrm{E}_{tl}^{\lambda}, R[b'_l; m], u_l)}{H}, & \textit{иначе} \end{cases}, \quad (23)$$

$$b_1 = 1, H = \sqrt{U^2 + m^2}, U = \sum_{\kappa=1}^{l} u_\kappa.$$

Благодаря возможности выполнить отказ от распознавания в случае невозможности найти границы всех образов $\mathrm{E}_{t_\kappa}^{t_{\kappa+1}}$ пар аллофонов из транскрипции $T = t_1, .., t_\kappa, ..., t_l$ в R, предложенная мера

расхождения $F(T,R)$ учитывает фонетический состав речевой команды (транскрипцию T) и тем самым решает проблему пропуска алгоритмом DTW отличающихся элементов и учёта сходных.

9. Разработка метода классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева на примере пофонемного распознавания с иерархическим представлением словаря классов речевых команд. Приведенный выше метод пофонемного распознавания применим только для распознавания речевых команд из небольшого словаря, так как вычисление функционала $F(T,R)$ выполняется отдельно для каждой транскрипции речевой команды.

Преобразуем словарь классов речевых команд W (17), заменив в нём представление множества фонетических транскрипций речевых команд $G=\{T_y\}$ на усовершенствованное trie-дерево (10):

$$W=\langle A,\mathrm{E},O,Ch,X\rangle. \tag{24}$$

Разработаем метод пофонемного распознавания, используя это иерархическое представление словаря классов речевых команд. Поставим в соответствие вершине $\mathbf{o}_j$ набор значений:

1) b_j – левая граница в R аллофона, которому соответствует вершина $\mathbf{o}_j$;

2) $dist_j$ – число, обозначающее найденную интегральную меру расхождения между фрагментом распознаваемой последовательностей $R[1;b_j]$ и начальной частью множества речевых команд, транскрипция которых хранится в ветви от корня дерева структуры до вершины $\mathbf{o}_j$;

3) len_j – суммарная длина эталонов цепочки аллофонов (в ветви от корня дерева до вершины $\mathbf{o}_j$), использованных при распознавании фрагмента ряда R, соответствующего этой цепочке.

$TW(k_i)$ – функция получения номера следующей вершины при обходе вершин дерева в глубину; $TW1(k_i)$ – функция получения номера следующей вершины при обходе вершин дерева вверх по ветви

до первого правого потомка. Функции $TW(k_i)$ и $TW1(k_i)$ возвращают номера вершин или 0.

$V = \{v_j\}_{j=1}^{z}$ – массив интегральных мер расхождения между распознаваемым рядом R и транскрипцией словаря речевых команд с идентификатором j.

Обозначим Y множество номеров вершин, в которых оканчиваются транскрипции распознанных команд.

Используем также промежуточные переменные tmp_d, u, Id.

Пофонемное распознавание R (1) выполняем путём обхода вершин (O, Ch, X) усовершенствованного trie-дерева (Алгоритм 2).

Пусть $N = |O|$ – количество вершин, а $z = |X|$ – количество последовательностей в дереве. Переменной k обозначим номер вершины, обрабатываемой на текущем шаге обхода дерева; вершины нумеруем с 1 ($\mathbf{o}_1$ – корневая вершина); значение 0 обозначает отсутствие искомой вершины в дереве. При инициализации: назначаем текущей вершиной корень дерева ($k := 1$); границей начала первого аллофона для всех последовательностей в дереве – 1-й вектор ряда ($b_k := 1$); очищаем множество Y; обнуляем b_j, $dist_j$, len_j, связанные с вершинами $\mathbf{o}_j$; обнуляем v_j, связанные с идентификаторами j последовательностей, хранимых в дереве. Выполняем обход вершин дерева, $\mathbf{o}_k$ – текущая вершина.

Возможна ситуация, когда вычисленная для вершины $\mathbf{o}_k$ граница начала аллофона больше или равна длине ряда R ($b_k \geq m$), или для вершины $\mathbf{o}_k$ произошел отказ от распознавания ($dist_k < 0$).

В этом случае последовательности образов аллофонов, начало которых соответствует фрагменту ветви от корня до вершины $\mathbf{o}_k$, не могут быть результатом распознавания ряда R – переходим к следующей вершине при обходе вверх по ветви до первого правого потомка.

Если в вершине оканчивается последовательность ($o_2^k \neq 0$), то вычисляем меру расхождения $\Phi'(E_g^{\lambda}, R[b_k; m], u)$ от образа $\mathrm{E}_{\mathrm{g}}^{\lambda}$

последнего аллофона g, которому соответствует вершина $\mathbf{o}_k$ (п. 2.2 Алгоритма 2), до оставшегося фрагмента ряда $R[b_k; m]$.

Если нет отказа от распознавания (п. 2.3.2 Алгоритма 2), то внести результаты сопоставления с R последовательности образов аллофонов с идентификатором o_2^k в массив V и множество Y.

1. *Инициализация*
   1.1 $N := |O|; z := |X|;\ k := 1; b_k := 1; Y := \varnothing;$
   1.2 *Для* $j = \overline{1,N}$ *выполнять* $b_j := 0; dist_j := 0; len_j := 0;$
   1.3 *Для* $j = \overline{1,z}$ *выполнять* $v_j := -1;$
2. *Пока* $(k \neq 0)$ *выполнять*

н.ц.

   2.1 *Если* $(b_k \geq m) \vee (dist_k < 0)$ *то* *Перейти на шаг* 2.7;
   2.2 $g := o_1^k;$
   2.3 *Если* $o_2^k \neq 0$ *то*
   2.3.1 $tmp\_d := \Phi'(\mathrm{E}_g^{\lambda}, R[b_k; m], u);$
   2.3.2 *Если* $(tmp\_d > -1)$
   то $Id := o_2^k; Y := Y \cup Id; v_{Id} := \dfrac{dist_k + tmp\_d}{\sqrt{(len_k + u)^2 + m^2}};$
   2.4 *Для* $n \in ch_{o_3^k}$ *выполнять*

н.ц.

   2.4.1 $h := o_1^n;$
   2.4.2 $tmp\_d := \Phi(\mathrm{E}_g^h, R[b_k; b_n], u);$
   2.4.3 *Если* $(tmp\_d > -1)$ $tmp\_d := -1;$
   то $dist_n := dist_k + tmp\_d; len_n := len_k + u;$
   2.4.4 *иначе* $b_n := m; dist_n := -1;$

к.ц.

   2.5 $k := TW(k);$
   2.6 *Если* $(k = 0)$
   то перейти на шаг 2.7;
   иначе перейти на шаг 2.8;
   2.7 $k := TW1(k);$
   2.8 ;

к.ц.

3. ***Результат:*** $j := \arg\min_{i \in Y}(v_i) : v_i \neq -1$

Алгоритм 2. Пофонемное распознавание с иерархическим представлением словаря классов речевых команд и нормировкой по длине эталона

Для всех вершин $\mathbf{o}_n$, являющихся потомками вершины $\mathbf{o}_k$, вычислить меру расхождения $\Phi(\mathrm{E}_g^h, R[b_k; b_n], u)$ до образа E_g^h где g и h соответствуют вершинам $\mathbf{o}_k$ и $\mathbf{o}_n$. Фиксируем результаты в массивах B, $Dist$, Len с учетом того, имел ли место отказ от распознавания.

Переходим к следующей вершине дерева.

Предложенный метод позволяет анализировать общие начальные части последовательностей образов (классов объектов) единожды, за счёт чего обеспечивает ускорение классификации последовательности векторов признаков на множестве последовательностей образов.

Численные исследования разработанного метода проведены на трёх словарях речевых команд, которые обозначим «С1», «С2», «С3». Перечень речевых команд, использованных для создания эталонов пар аллофонов, обозначен «Ф». В качестве речевых команд использованы слова русского языка, причем выбран набор слов, являющийся сложным для классификации алгоритмом DTW (слова фонетически близки; некоторые слова являются фрагментом других; слова существенно отличаются по длине). Перечни слов указанных словарей и слов, использованных при обучении эталонов пар аллофонов, приведен в таблице 1. Символ «\» обозначает ударение.

Результаты численных исследований приведены в таблицах 2-3. Распознаванию подвергались все слова словаря. Обучающая и тестовая выборка не пересекались.

Все слова произнесены одним диктором на одном звукозаписывающем оборудовании и в схожих акустических условиях. Для распознавания на основе алгоритма DTW каждая речевая команда имела один эталон, построенный по одной аудиозаписи произнесения речевой команды. Для пофонемного распознавания использованы эталоны пар аллофонов, которые построены по вручную размеченным аудиозаписям слов из набора «Ф», который приведен в таблице 1.

В таблице 2 приведено количество верно распознанных слов для словарей С1, С2, С3 при использовании определённого метода классификации. На рисунке 1 показано отношение числа верно распознанных команд к числу предъявленных для распознавания.

Из таблицы 2 видно, что и у ранее предложенного метода пофонемного распознавания [31] и у разработанного метода пофонемного распознавания с применением усовершенствованного trie-дерева совпадает количество верно распознанных команд. Это объясняется тем, что оба метода вычисляют меру расхождения по одному и тому же функционалу.

Таблица 1. Содержимое словарей С1, С2, С3 и перечень слов «Ф»

Слово	С3	С2	С1	Ф	Слово	С3	С2	С1	Ф	Слово	С3	С2	С1	Ф
\ил				+	м\ена	+				прот\ест	+			
\унт	+	+	+	+	м\есто	+	+	+	+	прот\опать	+	+	+	+
алл\ель	+	+			м\етка	+	+			с\ель	+			
б\оль	+	+	+	+	м\етод	+	+			с\ено	+			
багат\ель	+	+			мет\елица	+	+	+	+	с\ет	+			
бал\ет	+				мет\ель	+	+			с\етка	+			
бал\етки	+			+	мет\ил	+				с\етовать	+			
балаб\олить	+	+			минов\ать	+	+	+	+	своров\ать	+	+	+	+
балов\ать	+				набалов\ать	+	+	+	+	сел\ен	+			
вал\ет	+	+			наворов\ать	+				синаг\ога	+			
воров\ать	+	+			надков\ать	+				сов\ать	+			
г\оголь	+	+	+	+	надыш\ать	+	+	+	+	сол\ило	+			
г\огот	+				нал\ив	+				сос\ед	+			
г\оль	+	+			налив\ать	+				сос\едка	+	+	+	+
г\опать	+				налинов\ать	+	+	+	+	ст\оль	+			
гал\ета	+				налом\ать	+		+	+	став\ать	+			
гал\етка	+				нас\едка	+	+	+	+	стогов\ать	+	+	+	+
гогот\ун	+	+			нас\ест	+	+	+	+	т\ент	+	+		
д\ив	+	+			насев\ать	+	+	+	+	т\ест	+	+		
дыш\ать	+	+	+		нат\опать	+	+	+	+	т\есто	+	+		
заболев\ать	+	+			од\ышка	+	+		+	т\иф	+			
забалов\ать				+	однол\етка	+	+	+	+	т\ога	+			
завод\ила	+	+	+	+	ол\ифа	+	+	+	+	т\оль	+	+		
задыш\ать	+	+	+	+	от\ель	+	+			т\опать	+	+		
залив\ать	+	+			отков\ать	+	+			т\опливо	+			
засев\ать	+	+			пал\етка	+	+			т\ополь	+	+		
зат\опать	+	+	+		паралл\ель	+	+	+	+	т\опот	+	+		
кад\ило	+				паров\ать	+	+			т\ун	+	+		
калев\ать	+	+			пат\ент	+	+			уг\одно				+
карав\ай	+				патов\ать	+	+			улом\ать				+
карот\ель	+	+	+	+	пл\ед	+	+			ф\и	+	+		
кив\ать	+	+	+	+	пл\ен	+	+	+	+	ф\ила	+	+		
килев\ать	+	+	+	+	пл\и	+	+			ф\ифа	+	+	+	+
кис\ель	+	+			плев\ать		+			фас\ет	+	+		
кис\ет	+	+	+	+	плинтов\ать	+	+			фас\етка	+	+		
ков\ать	+				побалов\ать	+	+			хал\иф	+	+	+	+
кол\ено	+	+	+	+	поворов\ать	+	+			хохот\ун	+	+	+	+
кол\ет	+	+	+	+	под\и	+	+			целов\ать	+	+	+	+
колт\ун	+	+	+	+	подков\ать	+								
л\ай	+	+	+	+	подыш\ать	+								
л\ен	+	+			пол\ено	+								
л\ента	+	+	+	+	пол\ив	+								
л\ето	+	+			полив\ать	+								
л\етом	+	+	+	+	полинов\ать	+								
л\и				+	полом\ать	+								
л\иф	+				пос\етовать	+								
линов\ать	+	+			пот\опать	+								
лод\ыжка	+	+	+	+	продыш\ать	+	+	+	+					
лом\ать	+	+	+	+	прол\ив	+								
м\ай	+	+	+	+	пролив\ать	+								
м\ать	+	+			пролинов\ать	+								
м\елево	+	+	+	+	пролом\ать	+								
м\ель	+	+			просев\ать	+								

Таблица 2. Количество верно распознанных речевых команд

Словарь	Обозначение	C1	C2	C3
	Количество команд в словаре	45	91	138
Метод	Пофонемное распознавание	44	82	121
	Пофонемное распознавание с применением усовершенствованного trie-дерева	44	82	121
	DTW со взвешиванием по длине диагонали матрицы выравнивания	32	53	76
	DTW без взвешивания	32	51	74

В проведенном эксперименте у методов пофонемного распознавания отношение числа верно распознанных команд к числу предъявленных для распознавания более, чем на ¼ превзошло тот же показатель метода DTW. Это объясняется тем, что выбран словарь, демонстрирующий недостатки DTW (более подробно этот вопрос рассмотрен в работе [32]). Словари с такими характеристиками можно использовать для анализа работы методов распознавания в худшем случае, но нельзя использовать в реальных задачах распознавания речевых команд, так как речевые команды должны быть хорошо различимы.

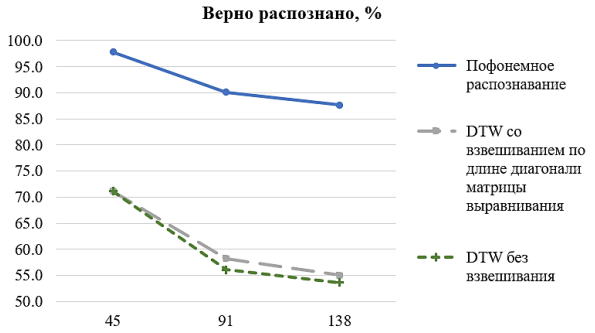


Рис. 1. Отношение числа верно распознанных команд к числу предъявленных для распознавания

Разработанный метод пофонемного распознавания с применением усовершенствованного trie-дерева отличается от предложенного ранее метод пофонемного распознавания [31] порядком, в котором вычисляют меры расхождения $\Phi(E_{t_к}^{t_{к+1}}, R[b'_к, b'_{к+1}], u_к)$ от образов пар аллофонов E множества

речевых команд до соответствующих фрагментов $R[b'_{\kappa}, b'_{\kappa+1}]$. Ранее предложенный метод сопоставлял с R каждую речевую команду отдельно. Разработанный метод общие начальные части речевых команд с соответствующими фрагментами R сопоставляет однократно, что снижает вычислительную сложность метода.

В таблице 3 показаны среднее время распознавания речевых команд в зависимости размера словаря и применённого метода.

Таблица 3. Среднее время распознавания речевых команд, мс

Словарь	Обозначение	C1	C2	C3
	Количество команд в словаре	45	91	138
Метод	Пофонемное распознавание	60	89	129
	Пофонемное распознавание с применением усовершенствованного trie-дерева	47	68	97
	DTW со взвешиванием по длине диагонали матрицы выравнивания	47	86	134
	DTW без взвешивания	47	86	137

На рисунке 2 отражено отношение времени распознавания различными методами ко времени распознавания разработанным методом.

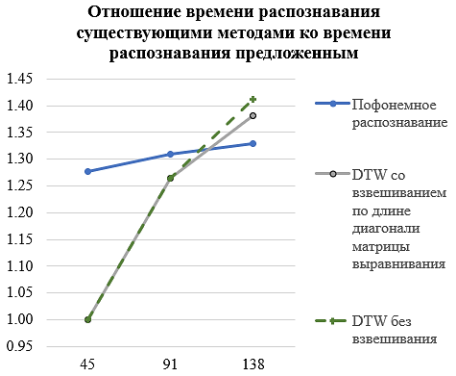


Рис. 2. Отношение времени распознавания различными методами ко времени распознавания разработанным методом

Для словаря C1 отношение времени распознавания методом DTW ко времени распознавания разработанным методом равно 1. Это можно объяснить тем, что количество речевого материала в эталонах метода DTW несколько меньше, чем в эталонах пар аллофонов,

используемых в пофонемном распознавании. Увеличение размера словаря влечёт увеличение количества речевого материала в эталонах метода DTW, что вызывает больший рост вычислительной сложности метода DTW, чем разработанного метода.

9. Заключение. Последовательности могут быть применены во многих областях. Прослеживается тенденция к обобщению задач из различных предметных областей как задач анализа последовательностей.

Впервые сформулированы две проблемы сильноветвящихся деревьев, которые решает усовершенствованное trie-дерево: проблема размера вершины, которая состоит в том, что вершины одного дерева имеют разный размер из-за разного числа вершин-потомков; использование для ссылки на вершины аппарата указателей. Практическая значимость решения этих проблем состоит в следующем. Фиксированный размер вершины и использование идентификаторов вершин вместо аппарата указателей позволяет: хранить и обрабатывать вершины дерева в структурах данных с произвольным доступом (например, файлах); обрабатывать деревья, частично загруженные в оперативную память. В проведенном эксперименте фиксированный размер вершины и использование идентификаторов вершин вместо аппарата указателей увеличило на 40% скорость полной загрузки из файла усовершенствованного trie-дерева. Выполнено формальное описание усовершенствованного trie-дерева с учетом решения проблем сильноветвящихся деревьев. Дано пояснение ранее полученным результатам о существенном приросте скорости операций добавления и удаления последовательностей в усовершенствованном trie-дереве относительно использования массива с индексной таблицей.

Разработан метод классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева на примере пофонемного распознавания с иерархическим представлением словаря классов речевых команд. Продемонстрировано, что усовершенствованное trie-дерево является способом представления множеств последовательностей, который можно эффективно применять в задачах классификации на множестве последовательностей элементов сложных объектов при использовании аддитивной или мультипликативной мер подобия / различия.

Численные исследования подтвердили, что метод пофонемного распознавания речевых команд как задачи классификации ряда на множестве последовательностей элементов сложных объектов решает

проблему пропуска алгоритмом DTW отличающихся элементов и учёта сходных.

Показано, что разработанный метод классификации ряда на множестве последовательностей элементов сложных объектов с применением усовершенствованного trie-дерева решает проблему вычислительной сложности алгоритма DTW за счёт однократного сопоставления общих начальных частей последовательностей с соответствующими им фрагментами ряда.

Пофонемное распознавание с иерархическим представлением словаря классов речевых команд не претендует на то, чтобы конкурировать с современными средствами распознавания речи. Пофонемное распознавание выбрано как «модельная задача» из хорошо знакомой автору предметной области. С помощью предложенного метода в будущем предполагается решать задачи классификации на множестве последовательностей элементов сложных объектов в других предметных областях для случая, когда количество размеченных данных недостаточно для обучения скрытой марковской модели или нейронной сети.

Литература

1. Вирт Н. Алгоритмы и структуры данных. Новая версия для Оберона + CD // М.: ДМК Пресс. 2010. 272 с.
2. Кнут Д.Э. Искусство программирования. Т.3: Сортировка и поиск // М.: Вильямс. 2000. 832 с.
3. Briandais R. File searching using variable-length keys // Proc. Western Joint Computer Conf. 1959. pp. 295–298.
4. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология // СПб.: Невский Диалект; БХВ-Петербург. 2003. 654 с.
5. Liao T.F., Bolano D., Brzinsky-Fay C., Cornwell B., Fasang A.E., Helske S., Piccarreta R., Raab M., Ritschard G., Struffolino E., Studer M. Sequence analysis: Its past, present, and future. Social science research. 2022. vol. 107. DOI: 10.1016/j.ssresearch.2022.102772.
6. Mathew S., Peat G., Parry E., Sokhal B.S., Yu D. Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review. Journal of Clinical Epidemiology. 2024. vol. 166. DOI: 10.1016/j.jclinepi.2023.111226.
7. Громов В.А., Мазайшвили К.В., Заикин П.В., Николаев Е.Н., Бесчастнов Ю.Н., Зворыкина Е.И., Паринов А.А., Незнанов А.А. Различение хаотических и регулярных временных рядов для идентификации состояния артериовенозной фистулы // Вестник кибернетики. 2022. № 1(45). С. 72–82.
8. Ковалева К.А., Яхонтова И.М. Теория исследования и разработки методов и моделей прогнозирования временных рядов с приращением в страховании // Новые технологии. 2019. № 4. С. 239–248.
9. Зюсько К.Д. Прогноз спроса на товар с помощью нейронных сетей в условиях меняющейся размерности входных данных // Экономика и качество систем связи. 2020. № 1 (15). С. 36–41.

10. Луценко Е.В. Применение автоматизированного системно-когнитивного анализа банковских баз данных по операциям с кредитными картами для количественной оценки риска мошенничества // Научный журнал КубГАУ. 2021. № 172. С. 82–172.
11. Кузьмин В.Н., Менисов А.Б. Исследование путей и способов повышения результативности выявления компьютерных атак на объекты критической информационной инфраструктуры // Информационно-управляющие системы. 2022. № 4. С. 29–43.
12. Leichtnam L., Totel E., Prigent N., Me L. Sec2graph: Network attack detection based on novelty detection on graph structured data // Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA. Springer International Publishing, 2020. pp. 238–258.
13. Жукова Н.А. Онтологические модели трансформации данных о состоянии технических объектов // Онтология проектирования. 2019. Т. 9. № 3(33). С. 345–360.
14. Nguyen D., Luo W., Nguyen T., Venkatesh S., Phung D. Sqn2Vec: Learning Sequence Representation via Sequential Patterns with a Gap Constraint. Machine Learning and Knowledge Discovery in Databases. Proceedings of the European Conference, ECML PKDD (Part II). 2019. pp. 569–584.
15. Fradkin D., Morchen F. Mining sequential patterns for classification. Knowledge and Information Systems. 2015. № 45 (3). pp. 731–749.
16. Привалов А.Н., Смирнов В.А. Метод нечеткого сравнения строк для обнаружения фейковых сайтов // Известия ТулГУ. Технические науки. 2022. № 2. С. 184–191.
17. Blanchard P. Sequence Analysis. Encyclopedia of Research Methods. London: Sage Publications Ltd. 2020. URL: https://www.researchgate.net/publication/342232021_Sequence_Analysis (дата обращения: 15.05.2024).
18. Vanasse A., Courteau J., Courteau M., Benigeri M., Chiu Y.M., Dufour I., Couillard S., Larivée P., Hudon C. Healthcare utilization after a first hospitalization for COPD: a new approach of State Sequence Analysis based on the '6W' multidimensional model of care trajectories. BMC Health Serv. Res. 2020. vol. 20(1). DOI: 10.1186/s12913-020-5030-0.
19. Su H., Liu S., Zheng B., Zhou X., Zheng K. A survey of trajectory distance measures and performance evaluation. The VLDB Journal. 2020. № 29. pp. 3–32.
20. Калихман И.Л., Войтенко М.А. Динамическое программирование в примерах и задачах: Учеб. Пособие. М.: Высш. школа, 1979. 125 с.
21. Коган Д.И. Динамическое программирование и дискретная многокритериальная оптимизация: учебное пособие. Нижний Новгород: Изд-во Нижегородского ун-та, 2004. 150 с.
22. Баширзаде Л.И., Алиев Г.С. Применение динамического программирования для моделирования процессов принятия решений // Архивариус. 2022. № 3 (66). С. 51–55.
23. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. К.: Наук. думка, 1987. 262 с.
24. Шелепов В.Ю., Дорохин О.А., Засыпкин А.В., Червин Н.А. О некоторых подходах к проблеме компьютерного распознавания устной русской речи // Труды Междунар. конф. «Знание – Диалог – Решение». 1997. Т. 1. С. 234–240.
25. Alshehri M., Coenen F., Dures K. Sub-sequence-based dynamic time warping. Proceedings of the 11th International Conference on Knowledge Discovery and Information Retrieval. 2019. pp. 274–281.

26. Deriso D., Boyd S. A general optimization framework for dynamic time warping // Optimization and Engineering. 2023. vol. 24. pp. 1411–1432.
27. Wang L., Koniusz P. Uncertainty-DTW for Time Series and Sequences. European Conference on Computer Vision (ECCV 2022). Cham: Springer Nature Switzerland. 2022. vol. 13681. pp. 176–195.
28. Bringmann K., Fischer N., Hoog I., Kipouridis E., Kociumaka T., Rotenberg E. Dynamic Time Warping // Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Publisher Society for Industrial and Applied Mathematics. 2024. pp. 208–242.
29. Jain V., Fokow V., Wicht J., Wetzker U. A Dynamic Time Warping Based Method to Synchronize Spectral and Protocol Domains for Troubleshooting Wireless Communication // IEEE Access. 2023. vol. 11. pp. 64668–64678.
30. Козлов А.В., Саввина Г.В., Шелепов В.Ю. Система пофонемного распознавания отдельно произносимых слов // Искусственный интеллект. 2003. № 1. С. 156–165.
31. Дорохина Г.В. Модификация алгоритма DTW для пофонемного распознавания слов // Проблемы искусственного интеллекта. 2015. № 0(1). С. 38–49.
32. Дорохина Г.В. Анализ методов распознавания речевых команд на основе алгоритма DTW // Труды шестого междисциплинарного семинара «Анализ разговорной русской речи» (АР3-2012) (27-28 августа 2012. г. Санкт-Петербург). 2012. С. 29–34.
33. Васильев В.И., Шевченко А.И., Эш С.Н. Принцип редукции в задачах обнаружения закономерностей: Монография. Донецк, 2009. 340 с.
34. Бурибаева А.К., Дорохина Г.В., Ниценко А.В., Шелепов В.Ю. Сегментация и дифонное распознавание речевых сигналов // Труды СПИИРАН. 2014. Т. 31. № 8. С. 20–42.
35. Дорохина Г.В., Павлюкова А.П. Модуль морфологического анализа слов русского языка // Искусственный интеллект. 2004. № 3. С. 636–642.
36. Дорохина Г.В. Патент на изобретение № UA 78806 «Устройство для хранения и поиска строковых величин и способ хранения и поиска строковых величин». собственник: Институт проблем искусственного интеллекта. Промышленная собственность. 2007. опубл. 25.04.2007.
37. Дорохина Г.В., Павлыш В.Н. Способ представления множеств последовательностей // Информатика и кибернетика. 2016. № 1(3). С. 56–64.
38. Дорохина Г.В. Сравнение затрат памяти для метода деревьев цифрового поиска и его усовершенствования // Искусственный интеллект. 2009. № 4. С. 338–343.
39. Финаев В.И., Дорохина Г.В. Применения усовершенствованных деревьев цифрового поиска // Проблемы искусственного интеллекта. 2019. № 4 (15). С. 62–77.
40. Bantay L., Abonyi J. Frequent pattern mining-based log file partition for process mining // Engineering Applications of Artificial Intelligence. 2023. vol. 123. DOI: 10.1016/j.engappai.2023.106221.
41. Xing Z., Pei J., Keogh J. A brief survey on sequence classification // SIGKDD Explor. 2010. vol. 12(1). pp. 40–48.
42. Atar R.H., Bhosale D.S. Pattern Based Sequence Classification // International Journal of Advanced Research in Science, Communication and Technology (IJARSCT). 2023. vol. 3. № 1. pp. 390–396.
43. Lazzari N., Poltronieri A., Presutti V. Classifying sequences by combining context-free grammars and OWL ontologies // European Semantic Web Conference. Cham: Springer Nature Switzerland, 2023. С. 156–173.
44. Crochemore M., Lecroq T, Liu L., Ozsu T. Encyclopedia of Database Systems. Verlag: Springer. 2009. pp. 3179–3182.

Дорохина Галина Владимировна — научный сотрудник, ФГБНУ "Институт проблем искусственного интеллекта". Область научных интересов: представление и обработка множеств последовательностей, динамическое программирование, распознавание образов, анализ текстов, разработка программного обеспечения, инженерия знаний, онтологии. Число научных публикаций — 48. sgv_iai@mail.ru; улица Артёма, 118б, 283048, Донецк, Россия; р.т.: +7(856)311-3424.

Поддержка исследований. Работа выполнена в рамках НИР №Г/Р 123092600030-4.

DOI 10.15622/ia.23.6.8

G. DOROKHINA

PHONEME-BY-PHONEME SPEECH RECOGNITION AS A CLASSIFICATION OF SERIES ON A SET OF SEQUENCES OF ELEMENTS OF COMPLEX OBJECTS USING AN IMPROVED TRIE-TREE

Dorokhina G. **Phoneme-by-Phoneme Speech Recognition as a Classification of Series on a Set of Sequences of Elements of Complex Objects Using an Improved Trie-Tree.**

Abstract. Sequences, including vector sequences, are applicable in any subject domains. Sequences of scalar values or vectors (series) can be produced by higher-order sequences, for example: a series of states, or elements of complex objects. This academic paper is devoted to the application of an improved trie-tree in the classification of series on a set of sequences of elements of complex objects using the dynamic programming method. The implementation areas of dynamic programming have been considered. It has been shown that dynamic programming is adapted to multi-step operations of calculating additive (multiplicative) similarity/difference measures. It is argued that the improved trie-tree is applicable in the problem of classifying a series on a set of sequences of elements of complex objects using such similarity/difference measures. An analysis of hierarchical representations of sets of sequences has been performed. The advantages of the improved trie-tree over traditional representations of other highly branching trees have been described. A formal description of the improved trie-tree has been developed. An explanation has been given to the previously obtained data on a significant speed gain for operations of adding and deleting sequences in the improved trie-tree relative to the use of an array with an index table (24 and 380 times, respectively). The problem of phoneme-by-phoneme recognition of speech commands has been formulated as a problem of classifying series on a set of sequences of elements of complex objects and a method for its solving has been presented. A method for classifying a series on a set of sequences of elements of complex objects using the improved trie-tree is developed. The method has been studied using the example of phoneme-by-phoneme recognition with a hierarchical representation of the dictionary of speech command classes. In this method, recognition of speech commands is executed traversing the improved trie-tree that stores a set of transcriptions of speech commands – sequences of transcription symbols that denote classes of sounds. Numerical studies have shown that classifying a series as sequences of elements of complex objects increases the frequency of correct classification compared to classifying a series on a set of series, and using the improved trie-tree reduces the time spent on classification.

Keywords: trie-tree, sets of sequences, classification of series on a set of sequences of elements of complex objects, dynamical programming, phoneme-by-phoneme recognition of speech commands.

References

1. Wirth N. Algorithms and Data Structures. Oberon version. 2004. 211 p.
2. Knut D.Je. Iskusstvo programmirovanija. Vol. 3: Sortirovka i poisk. [The Art of Computer Programming. Vol. 3: Sorting and Searching]. M.: Vil'jams. 2000. 832 p. (In Russ.).
3. Briandais R. File searching using variable-length keys. Proc. Western Joint Computer Conf. 1959. pp. 295–298.
4. Gusfield D. Algorithms on Strings, Trees, and Sequences – Computer Science and Computational Biology. Davis: University of California, 1997. 556 p.

5. Liao T.F., Bolano D., Brzinsky-Fay C., Cornwell B., Fasang A.E., Helske S., Piccarreta R., Raab M., Ritschard G., Struffolino E., Studer M. Sequence analysis: Its past, present, and future. Social science research. 2022. vol. 107. DOI: 10.1016/j.ssresearch.2022.102772.
6. Mathew S., Peat G., Parry E., Sokhal B.S., Yu D. Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review. Journal of Clinical Epidemiology. 2024. vol. 166. DOI: 10.1016/j.jclinepi.2023.111226.
7. Gromov V.A., Mazayshvili K.V., Zaikin P.V., Nikolaev E.N., Beschastnov Yu.N., Zvorykina E.I., Parinov A.A., Neznanov A.A. [Differentiating Chaotic and Regular Time Series for Identification of Arteriovenous Fistula State]. Vestnik kibernetiki – Proceedings in Cybernetics. 2022. no. 1(45). pp. 72–82. (In Russ.).
8. Kovaleva K.A., Yahontova I.M. [Research and development theory methods and models for forecasting time series with insurance increments]. Novye Tehnologii – New technologies. 2019. no. 4(50). pp. 239–248. (In Russ.).
9. Zyus`ko K.D. [Forecasting demand for goods using neural networks in conditions of changing dimensionality of input data]. E`konomika i kachestvo sistem svyazi – Economics and quality of communication systems. 2020. no. 1(15). pp. 36–41. (In Russ.).
10. Lucenko E.V. [Application of automated system-cognitive analysis of bank databases on credit card transactions to quantify the risk of fraud]. Nauchny`j zhurnal KubGAU – Scientific Journal of KubSAU. 2021. vol. 172. pp. 82–172. (In Russ.).
11. Kuz`min V.N., Menisov A.B. Investigation of ways and means to improve the effectiveness of detecting computer attacks on critical information infrastructure facilities. Informacionno-upravlyayushhie sistemy` – Information management systems. 2022. no. 4. pp. 29–43. (In Russ.).
12. Leichtnam L., Totel E., Prigent N., Me L. Sec2graph: Network attack detection based on novelty detection on graph structured data. Detection of Intrusions and Malware, and Vulnerability Assessment: 17th International Conference, DIMVA. Springer International Publishing, 2020. pp. 238–258.
13. Zhukova N.A. [Ontological models of transformation of data on the state of technical objects]. Ontologiya proektirovaniya – Design Ontology. 2019. vol. 9. no. 3(33). pp. 345–360. (In Russ.).
14. Nguyen D., Luo W., Nguyen T., Venkatesh S., Phung D. Sqn2Vec: Learning Sequence Representation via Sequential Patterns with a Gap Constraint. Machine Learning and Knowledge Discovery in Databases. Proceedings of the European Conference, ECML PKDD (Part II). 2019. pp. 569–584.
15. Fradkin D., Morchen F. Mining sequential patterns for classification. Knowledge and Information Systems. 2015. № 45 (3). pp. 731–749.
16. Privalov A.N., Smirnov V.A. [Fuzzy string match method for detecting fake sites]. Izvestiya TulGU. Tehnicheskie nauki – News of the Tula state university. Technical sciences. 2022. no. 2. pp. 184–191. (In Russ.).
17. Blanchard P. Sequence Analysis. Encyclopedia of Research Methods. London: Sage Publications Ltd. 2020. URL: https://www.researchgate.net/publication/342232021_Sequence_Analysis (дата обращения: 15.05.2024).
18. Vanasse A., Courteau J., Courteau M., Benigeri M., Chiu Y.M., Dufour I., Couillard S., Larivée P., Hudon C. Healthcare utilization after a first hospitalization for COPD: a new approach of State Sequence Analysis based on the '6W' multidimensional model of care trajectories. BMC Health Serv. Res. 2020. vol. 20(1). DOI: 10.1186/s12913-020-5030-0.

19. Su H., Liu S., Zheng B., Zhou X., Zheng K. A survey of trajectory distance measures and performance evaluation. The VLDB Journal. 2020. № 29. pp. 3–32.
20. Kalihman I.L., Vojtenko M.A. Dinamicheskoe programmirovanie v primerah i zadachah: Ucheb. Posobie [Dynamic programming in examples and problems: Textbook]. Moskow: Vyssh. shkola, 1979. 125 p. (In Russ.).
21. Коган Д.И. Динамическое программирование и дискретная многокритериальная оптимизация: учебное пособие. Нижний Новгород: Изд-во Нижегородского ун-та, 2004. 150 с.
22. Bashirzade L.I., Aliev G.S. [Application of dynamic programming to modeling decision making processes]. Arhivarius. 2022. no. 3(66). pp. 51–55. (In Russ.).
23. Vintsyuk T.K. Analiz, raspoznavaniye i interpretatsiya rechevykh signalov [Analysis, recognition and interpretation of speech signals]. K.: Nauk. dumka, 1987. 262 p. (In Russ.).
24. Shelepov V.Y., Dorokhin O.A., Zasipkin A.V., Chervin N.A. [On some approaches to the problem of computer speech recognition of spoken Russian] «Znanie – Dialog – Reshenie»: trudy Mezhdunar. konf. [Proceedings of the Intern. Conf. "Knowledge – Dialogue – Solution"]. 1997. vol. 1. pp. 234–240. (In Russ.).
25. Alshehri M., Coenen F., Dures K. Sub-sequence-based dynamic time warping. Proceedings of the 11th International Conference on Knowledge Discovery and Information Retrieval. 2019. pp. 274–281.
26. Deriso D., Boyd S. A general optimization framework for dynamic time warping. Optimization and Engineering. 2023. vol. 24. pp. 1411–1432.
27. Wang L., Koniusz P. Uncertainty-DTW for Time Series and Sequences. European Conference on Computer Vision (ECCV 2022). Cham: Springer Nature Switzerland. 2022. vol. 13681. pp. 176–195.
28. Bringmann K., Fischer N., Hoog I., Kipouridis E., Kociumaka T., Rotenberg E. Dynamic Time Warping. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Publisher Society for Industrial and Applied Mathematics. 2024. pp. 208–242.
29. Jain V., Fokow V., Wicht J., Wetzker U. A Dynamic Time Warping Based Method to Synchronize Spectral and Protocol Domains for Troubleshooting Wireless Communication. IEEE Access. 2023. vol. 11. pp. 64668–64678.
30. Kozlov A.V., Savvina G.V., Shelepov V.U. [Isolated word recognition system based on phoneme recognition]. Iskusstvennyj intellekt – Artificial Intelligence. 2003. vol. 1. pp. 156–165. (In Russ.).
31. Dorokhina G.V. [Modification of the algorithm DTW for spoken word recognition based on phoneme recognition]. Problemy iskusstvennogo intellekta – Problems of artificial intelligence. 2015. vol. 0(1). pp. 38–49. (In Russ.).
32. Dorokhina G.V. [Analysis of speech command recognition methods based on the DTW algorithm] Trudy` shestogo mezhdisciplinarnogo seminara «Analiz razgovornoj russkoj rechi» (AR3-2012) [Proceedings of the sixth interdisciplinary seminar «Analysis of Spoken Russian Speech» (AR3-2012)]. 2012. pp. 29–34. (In Russ.).
33. Vasil'yev V.I., Shevchenko A.I., Esh S.N. Printsip reduktsii v zadachakh obnaruzheniya zakonomernostey: Monografiya [The principle of reduction in the problems of detecting patterns: Monograph]. Donetsk, 2009. 340 p. (In Russ.).
34. Buribayeva A.K., Dorokhina G.V., Nitsenko A.V., Shelepov V.Ju. [Segmentation and diphone recognition of speech signals]. Trudy SPIIRAN – SPIIRAS Proceedings. 2014. vol. 31. no. 8. pp. 20–42. (In Russ.).
35. Dorokhina G.V., Pavlyukova A.P. [Module of morphological analysis of words of the Russian language]. Iskusstvennyy intellect – Artificial Intelligence. 2004. № 3. pp. 636–642. (In Russ.).

36. Dorokhina G.V. Patent No. UA 78806 “Device for saving and searching for lowercase values and method for saving and searching for lowercase values” Owner: Institute of problems of artificial intelligence Promyshlennaja sobstvennost' [Industrial property]. 25.04.2007. (In Russ.).
37. Dorokhina G.V., Pavlysh V.N. [A method of presenting sets of sequences]. Informatika i kibernetika – Informatics and Cybernetics. 2016. № 1(3). pp. 56–64. (In Russ.).
38. Dorokhina G.V. [Memory expenses comparison for the method of digital search tree and its improvement]. Iskusstvennyj intellekt – Artificial Intelligence. 2009. vol. 4. pp. 338–343. (In Russ.).
39. Finayev V.I., Dorokhina G.V. [Applications of improved digital search trees]. Problemy iskusstvennogo intellekta – Problems of artificial intelligence. 2019. vol. 4 (15). pp. 62–77. (In Russ.).
40. Bantay L., Abonyi J. Frequent pattern mining-based log file partition for process mining. Engineering Applications of Artificial Intelligence. 2023. vol. 123. DOI: 10.1016/j.engappai.2023.106221.
41. Xing Z., Pei J., Keogh J. A brief survey on sequence classification. SIGKDD Explor. 2010. vol. 12(1). pp. 40–48.
42. Atar R.H., Bhosale D.S. Pattern Based Sequence Classification. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT). 2023. vol. 3. № 1. pp. 390–396.
43. Lazzari N., Poltronieri A., Presutti V. Classifying sequences by combining context-free grammars and OWL ontologies. European Semantic Web Conference. Cham: Springer Nature Switzerland, 2023. C. 156–173.
44. Crochemore M., Lecroq T, Liu L., Ozsu T. Encyclopedia of Database Systems. Verlag: Springer. 2009. pp. 3179–3182.

Dorokhina Galina — Research fellow, FSBSI “IPAI. Research interests: representation and processing of sets of sequences, dynamic programming, pattern recognition, text analysis, software development, knowledge engineering, ontology. The number of publications — 48. sgv_iai@mail.ru; 118b, Artem St., 283048, Donetsk, Russia; office phone: +7(856)311-3424.

Acknowledgements. The work was carried out within the research No. G/R 123092600030-4.

DOI 10.15622/ia.23.6.9

M. ELLAKKIYA, T. RAVI, S. PANNEER AROKIARAJ

RUZICKA INDEXIVE THROTTLED DEEP NEURAL LEARNING FOR RESOURCE-EFFICIENT LOAD BALANCING IN A CLOUD ENVIRONMENT

Ellakkiya M., Ravi T., Panneer Arokiaraj S. **Ruzicka Indexive Throttled Deep Neural Learning for Resource-Efficient Load Balancing in a Cloud Environment.**

Abstract. Cloud Computing (CC) is a prominent technology that permits users as well as organizations to access services based on their requirements. This computing method presents storage, deployment platforms, as well as suitable access to web services over the internet. Load balancing is a crucial factor for optimizing computing and storage. It aims to dispense workload across every virtual machine in a reasonable manner. Several load balancing techniques have been conventionally developed and are available in the literature. However, achieving efficient load balancing with minimal makespan and improved throughput remains a challenging issue. To enhance load balancing efficiency, a novel technique called Ruzicka Indexive Throttle Load Balanced Deep Neural Learning (RITLBDNL) is designed. The primary objective of RITLBDNL is to enhance throughput and minimize the makespan in the cloud. In the RITLBDNL technique, a deep neural learning model contains one input layer, two hidden layers, as well as one output layer to enhance load balancing performance. In the input layer, the number of cloud user tasks is collected and sent to hidden layer 1. In that layer, the load balancer in the cloud server analyzes the virtual machine resource status depending on energy, bandwidth, memory, and CPU using the Ruzicka Similarity Index. Then, it is classified VMs as overloaded, less loaded, or balanced. The analysis results are then transmitted to hidden layer 2, where Throttled Load Balancing is performed to dispense the workload of weighty loaded virtual machines to minimum loaded ones. The cloud server efficiently balances the workload between the virtual machines in higher throughput and lower response time and makespan for handling a huge number of incoming tasks. To evaluate experiments, the proposed technique is compared with other existing load balancing methods. The result shows that the proposed RITLBDNL provides better performance of higher load balancing efficiency of 7%, throughput of 46% lesser makespan of 41%, and response time of 28% than compared to conventional methods.

Keywords: cloud computing, load balancing, deep learning, Ruzicka similarity index, throttled load balancing.

1. Introduction. CC is a paradigm that includes distributing services and resources more than the internet. Load balancing (LB) in CC is a significant aspect that is a pivotal task in optimizing resource utilization, enhancing performance, and ensuring high availability of applications and services. As cloud environments consist of multiple servers and resources, distributing incoming user requests efficiently among virtual machines becomes essential to prevent the overloading of any single machine in the cloud server. This distribution of workload helps in achieving optimal resource consumption, enhancing the efficiency of applications.

Task Scheduling- DT (TS-DT) method was developed [1] to distribute and execute tasks within an application. The algorithm

successfully enhances load balancing and reduces makespan but it failed in achieving energy-aware load balancing with minimal response time. A P2BED-C was developed in [2] to minimize energy consumption. However, the efficiency of the method was not improved.

A Reinforcement Learning (RL) model was developed in [3] to optimize cloud resource utilization for providing the best Quality of Service (QoS). However, the makespan was not efficiently reduced. A Dynamic and Resource-Aware Load Balancing technique was introduced [4] to enhance throughput and reduce makespan. However, a resource-aware scheduling approach was not employed for the distribution of tasks on virtual machines (VMs).

The Predictive Priority-based Modified Heterogeneous algorithm was designed in [5] to achieve efficient and dynamic resource provisioning for end user's requirements. However, it did not implement a more effective resource provisioning scheme for end-users. The Bio-Inspired Improved Lion Optimization method was designed in [6], to address load balancing issues through enhancing throughput as well as reducing migration time. However, the performance of efficiency remained unaddressed.

A content-aware machine learning technique was introduced in [7] for enhancing load balancing, leading to improved throughput and minimized response time. However, failed to reduce migration time. Dynamic load balancing method was developed in [8] by Q-learning for resource allocation, resource accessibility, and consideration of user preferences with the aim of minimizing response time and resource consumption. However, it did not achieve higher efficiency in a multitasking environment.

In [9], a multi-objective task scheduling technique was designed with the aim of optimizing scheduling, increasing throughput, as well as reducing both makespan and resource utilization. However, it did not address the minimization of response time. A dynamic virtual machine consolidation method was introduced in [10] for LB to mitigate tradeoffs among energy utilization as well as time complexity.

1.1. Contributions in this article are as follows. The main contributions of the paper as given below.

To enhance load balancing efficiency, the RITLBDNL technique has been developed by Deep Neural Learning and Throttled Load Balancing.

The RITLBDNL technique utilizes the Ruzicka Similarity Index to analyze incoming user tasks and determine the resource status of VMs.

The Throttled Load Balancing process is applied to deep neural learning for task migration from heavily loaded virtual machines to less loaded virtual machines with higher efficiency.

Finally, comprehensive and comparative experiments have been conducted to perform quantitative analysis using various performance metrics.

1.2. Paper organization. The remainder of this article is organized into dissimilar sections: Section 2 explains the literature survey. Section 3 presents the RITLBDNL method. Section 4 details the experimental analysis and describes the dataset. Section 5 gives a performance assessment of the proposed algorithm in comparison to conventional techniques. At last, section 6 gives conclusions of the paper.

2. Literature survey. Load Balancing Protocol was developed in [11] for CC with the aim of minimizing Makespan as well as throughput of VM utilization. Long Short-Term Memory Networks (LSTM) Machine Learning (ML) algorithm was designed in [12] for enhancing load balancing through optimized resource allocation. However, it did not succeed in enhancing the system performance of LB. An integrated optimization algorithm was developed in [13] to make an effective load balancing system that guarantees resource utilization and minimizes task response time.

Component-based throttled load balancing method was introduced in [14], but it failed to consider additional parameters for ensuring the optimal performance of load balancing algorithms. The Receiver-Initiated Deadline-Aware LB approach was developed in [15], and aimed to facilitate migration of incoming tasks to suitable virtual machines. However, this approach was not employed for scientific workflow applications for diverse QoS parameters.

An Action-Based Load Balancing scheme was designed [16] with the aim of reducing makespan and optimizing resource utilization. However, it failed to address resource allocation and management concepts within a cloud data center. A new resource optimization framework was introduced in [17] specifically designed for achieving load balancing with minimal resource utilization. An optimal load balancing method was developed [18], which effectively balances the load on cloud servers.

A re-modified throttled algorithm was developed in [19] to minimize the risk of load imbalance by considering the availability of VMs. However, it failed to address the issues related to increasing the efficiency of the algorithm. A load balancing approach based on renewable energy was developed in [20] to optimize interactive task allocation, aiming to reduce energy costs.

Modified honeybee behavior load balancing (HBB-LB) was introduced in [21] to secure the cloud system. However, the system performance was not enhanced. The Sine Cosine-based Elephant Herding

Optimization (SCEHO) algorithm was combined in [22] by Improved Particle Swarm Optimization (IPSO). Task scheduling behavior was improved but, throughput was not increased.

The two-stage genetic mechanism was utilized in [23] to monitor and manage VMH. But, it failed to minimize the time. A deep load balancer was introduced in [24] to allocate resources with less delay. Nevertheless, it failed to enhance throughput. Improved Lion Optimization (ILO) with Min-Max Algorithm was developed in [25] to identify the optimum solution. However, the load balancing efficiency was not sufficient.

3. Proposal methodology. In cloud computing, dynamically provisioning the resources for applications is a key and challenging task. However, cloud providers face resource management concerns due to inconsistent workloads in heterogeneous environments. The cloud service provider focuses on resource consumption, while the cloud user aims to achieve a shorter makespan time. Therefore, achieving load balancing is a significant parameter for effective task execution to obtain optimal consumption of cloud resources. A new RITLBDNL method is developed for efficient load balancing in a cloud computing environment. Figure 1 depicts a diagram of the RITLBDNL method for efficient LB in the cloud.

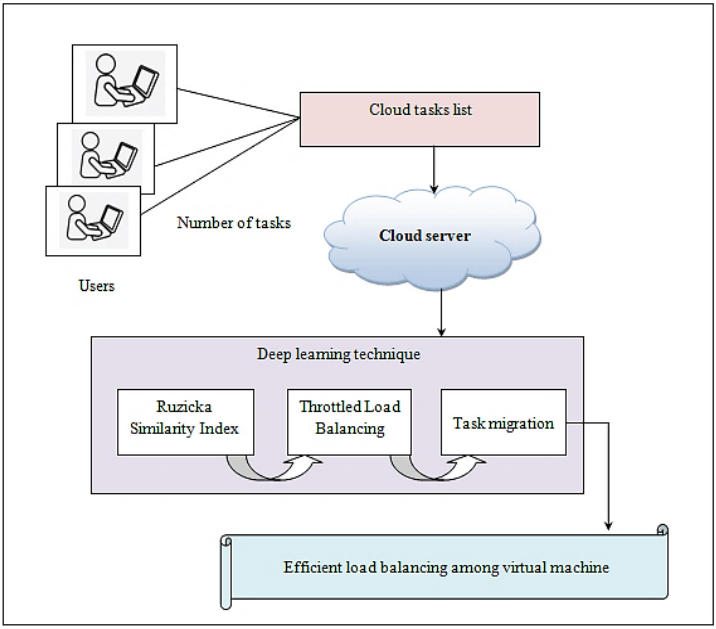


Fig. 1. Architecture diagram of the proposed RITLBDNL technique

Figure 1 demonstrates the RITLBDNL technique uses the deep learning concept for efficient load balancing in the cloud. The four components as cloud user (U), cloud server (CS), load balancer (LB), and virtual machines (V_m) included in the above figure. The working mechanism of the RITLBDNL model uses deep neural learning with several layers. The technique collects the number of cloud user requests or tasks. Ruzicka Similarity Index is utilized in hidden layer 1 to examine the virtual machine resource status. In hidden layer 2, the workload from heavily loaded virtual machines to less loaded ones is distributed to perform task migration by Throttled Load Balancing. In this way, throughput is improved and response time and makespan are minimized.

3.1. System model. It involves four key entities namely cloud user (U), cloud server (CS), load balancer (LB), and virtual machines (V_m). Initially, the cloud user 'U' submits numerous tasks, denoted as $T = \{T_1, T_2, \dots, T_n\}$, to the cloud server (CS). CS receives these tasks as of U. Subsequently, the load balancer within the cloud server analyzes and determines the status of virtual machines, categorizing them as minimum loaded, overloaded, as well as balanced load capacity. Once VM statuses are identified, the load balancer executes task migration using throttled load balancing with higher efficiency.

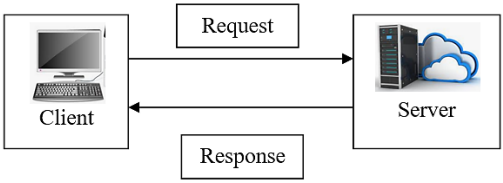


Fig. 2. Overview model of the client-server system

In above Figure 2, the client-server model includes the Server or Client. The client-server model explains the communication among two computing entities over a network. A client is a program that creates requests to a server. A server is a program that responds to those requests.

3.2. Ruzicka Indexive Throttle Load Balanced Deep Neural Learning. Deep Learning (DL) is a type of ML, which focuses on the development, and training of Artificial Neural Network (ANN) to perform some process. The term "deep" refers to the use of Deep Neural Networks (DNNs), which contain numerous hidden layers among input as well as output layers. These networks are referred to as DNNs or DL models.

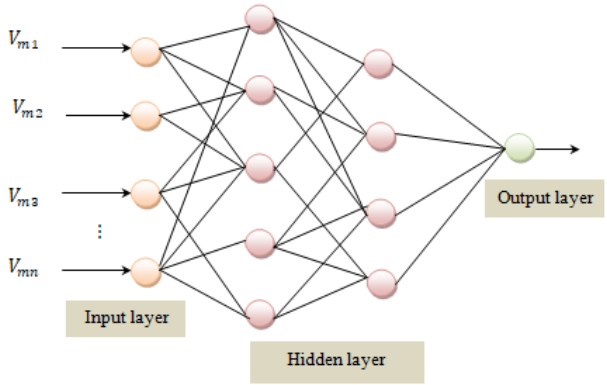


Fig. 3. Structures of the deep neural networks

Figure 3 depicts the structure of DNNs. The DNN is a fully connected feed-forward artificial neural network and it generates a set of outputs from a set of inputs. A DNN is constructed with three main layers such as input, hidden (i.e. middle), and output layers. The input and output layers are always single layers, whereas the middle layer includes two sublayers for analyzing the given input. Each layer is typically composed of small individual units called artificial neurons or nodes. The artificial neuron has the ability to process the given weighted inputs and applies an activation function and forward output to other nodes in the network. An input to an artificial neuron is a number of virtual machines (Vm_i). The neuron in one layer is fully connected to the neuron in another layer.

Each connection between neurons has an associated weight, which determines the strength of the connection. It also has an associated bias. The equation for a single neuron is expressed mathematically as follows:

$$Y = \mathrm{F}[\mathrm{S}], \tag{1}$$

$$\mathrm{S}=\sum_{i=1}^{n} (Vm_i * w_{ij}) + Q, \tag{2}$$

where Y indicates an output of the neuron, w_{ij} denotes the weight of the connection between the i^{th} neuron in the previous layer and the j^{th} neuron in the current layer, $Vm_i * w_{ij}$ denotes a product of the weight (w_{ij}) associated with the connection between the i^{th} neuron in the previous layer and j^{th} neuron in the current layer, and the input (Vm_i i.e $virtual\ machine$) from the i^{th} neuron in the previous layer. From

(2), 'Q' indicates a bias term that stores the numeric value as one, F denotes a sigmoid activation function used to determine whether a neuron is activated or not, it suggests that the neuron output is binary, typically representing a binary classification decision (activated or not activated).

$$\mathrm{F}\,[\mathrm{S}] = \frac{1}{1+\exp(-\mathrm{S})}, \tag{3}$$

where F $[S]$neuron's output with sigmoid activation is passed to the next layer of neurons

The input is transferred to a hidden layer where the resource availability of a virtual machine is determined to facilitate efficient load balancing.

$$RA\,(Vm_i) = \{Mem_c, BaW_C, E_C, CPU_{Ut}\}, \tag{4}$$

where $RA\,(Vm_i)$ denotes a resource's availability of the virtual machine that includes a memory capacity 'Mem_c', bandwidth capacity 'BaW_C', energy capacity 'E_C' and CPU utilization 'CPU_{Ut}'.

Initially, the memory capacity is determined by calculating the variance among total available memory as well as utilized memory.

$$Mem_c = To_{Memc} - Con_{Memc}, \tag{5}$$

where Mem_c indicates the memory capacity of the VM and To_{Memc} indicates the total memory capacity, Con_{Memc} represents the utilized memory capacity. Variation between the total memory capacity and the utilized memory capacity measurement is employed to assess the present memory status of the VM.

The bandwidth of a virtual machine denotes its capability to handle the maximum amount of data, typically measured in Mbps (megabits per second). The current status of bandwidth is determined through mathematical calculations.

$$BaW_C = Ba_{VM\,(total)} - Ba_{VM\,(con)}, \tag{6}$$

where BaW_C denotes the bandwidth capacity, $Ba_{VM\,(total)}$ indicates the total bandwidth, and $Ba_{VM\,(con)}$ represents the utilized bandwidth. Depending on the above-said metrics, the current status of the bandwidth capacity is determined.

The total energy consumption is calculated depending on the energy usage of the VM. Energy utilization is measured in kWh. Thus, the energy capacity of the virtual machine is determined as follows:

$$E_C = [Tot_E] - [Con_E], \tag{7}$$

where E_C represents the energy capacity, Tot_E indicates the total energy, Con_E denotes the consumed energy.

The CPU utilization time of the VM is computed mathematically by calculating the variance between the total time and the time spent processing specific tasks. This calculation helps to assess the efficiency and resource consumption of the virtual machine during the execution of its assigned workload.

$$CPU_{Ut} = [T_{cpu}] - [C_{cpu}], \tag{8}$$

where CPU_{Ut} denotes the CPU time, T_{cpu} indicates the total time and C_{cpu}symbolizes the consumed time of VM.

The proposed RITLBDNL technique finds the resource availability of a virtual machine based on the energy, bandwidth, memory, and CPU through the similarity measure. Ruzicka Similarity Index is employed for discovering the similarity between two sets. It provides a range from 0 to 1. Ruzicka Similarity Index is used to analyze the VM resource status as well as categorize VM as OL, minimum loaded and BL. The mathematical formula for calculating the similarity between the nodes is shown below

$$\beta = \frac{[RA\ (Vm_i)\ \cap\ T)\ ]}{\sum RA\ (Vm_i)\ + \sum T - [RA\ (Vm_i)\ \cap\ T)\ ]}, \tag{9}$$

where 'β' denotes a Ruzicka similarity coefficient, $RA\ (Vm_i)$ denotes the resource availability of the virtual machine and T indicates the threshold (i.e., 0.5),$RA\ (Vm_i)\ \cap\ T$ denotes a mutual dependence between the resource availability and threshold. The coefficient (β) provides the output ranges between 0 and 1. Likewise, similarities of all the VMs are computed based on the energy, bandwidth, and memory, and CPU using the statistic similarity coefficient

$$\beta = \begin{cases} < 0.5,\ UL \\ = 0.5\ ,\ BL, \\ > 0.5,\ OL \end{cases} \tag{10}$$

where β denotes the output of coefficient. Depend on coefficient outcome, LB determines over loaded (OL) , under loaded (UL) and balanced load (BL).

Throttled Load Balancing refers to a type of load balancing mechanism that includes throttling. Load balancing is the process of dispensing network traffic or calculating workload across numerous resources to guarantee no one, virtual node is overloaded. Throttling, in this context, involves controlling the rate at which certain requests are processed to manage the load on the system.

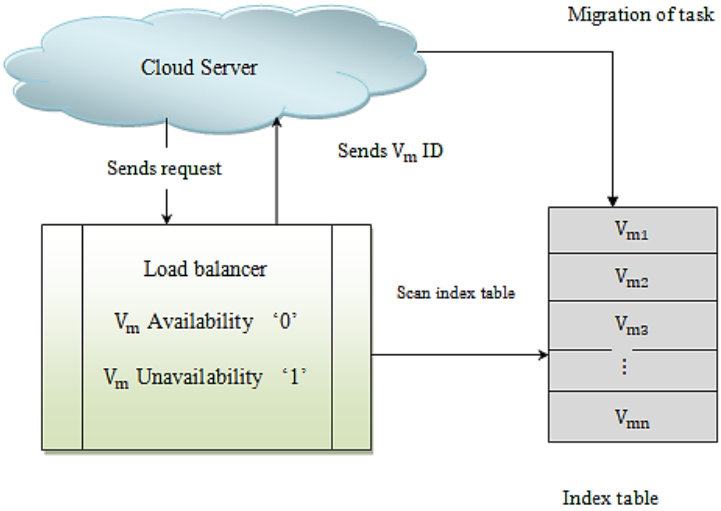


Fig. 4. Flow Process of throttled load balancing

Figure 4 depicts the flow process of throttled load balancing that contains cloud server, LB and several V_m $V_{m1}, V_{m2}, V_{m3}, \ldots V_{mn}$. Initially, a number of tasks are sent to CS. Then the server sends requests to the load balancer to identify the accessibility of the VM.

$$CS \overset{req}{\Longrightarrow} LB, \tag{11}$$

where req denotes the request. After receiving the request, the load balancer maintains a complete list of virtual machines using an index table and responds with the availability status.

$$\text{LB} \rightarrow \begin{cases} 1;\ V_m'sunavailable \\ 0;\ V_m'savailable \end{cases}, \tag{12}$$

where the status of $V_m's$ is identified through '1' and '0'.After that, the LB starts to scan index tables and send the less-loaded and heavily loaded VM IDs' to the cloud server. The server performs tasks migration from a heavily loaded to a less loaded virtual machine. In this way, resource-efficient load balancing is obtained at the output layer. The algorithm of Ruzicka indexive Throttle Load Balanced Deep neural learning is given below.

//Algorithm 1: Ruzicka indexive Throttle Load Balanced Deep neural learning

Input: Number of cloud user requests $T_1, T_2, T_3, \ldots, T_n$; virtual machines $(Vm_1, Vm_2, \ldots, Vm_n)$; cloud server (CS), loads balancer (LB),

Output: Increase the load balancing efficiency

Begin

1. Send number of requests or tasks $T_1, T_2, T_3, \ldots, T_n$ to CS
2. LB find the resoure capcity of virtual machine 'Vm' **—hidden layer 1**
3. **For each** virtual machine 'Vm'
4. Compute the multiple resources '$RA\ (Vm_i)$' using (5) (6) (7) (8)
5. Measure the Ruzicka similarity coefficient 'β'
6. ***if*** $(\beta > 0.5)$ **then**
7. virtual machine is classified as a overloaded
8. ***elseif*** $(\beta = 0.5)$ **then**
9. virtual machine is classified as a balanced loaded
10. ***elseif*** $(\beta < 0.5)$ **then**
11. virtual machine is classified as a less loaded
12. **End if**
13. CS send request to LB **---hidden layer**
14. LB sends virtual machine availability status to server
15. LB scan the index table
16. LB sends the less loaded and overloaded virtual machine ID's to server
17. CS perform task migration from overloaded to less loaded virtual machine
18. Obtain final load balancing **at the output layer**
19. **End for**

End

Algorithm 1 described above outlines the process of load balancing by the Ruzicka Indexive Throttle Load-Balanced Deep Neural Learning approach. For each incoming task from the user, the load balancer in the cloud server estimates the resource availability of the VM by Ruzicka Index function. This function is utilized to calculate the load status of each VM in the first hidden layer, classifying them as less loaded, overloaded, and balanced loaded. Subsequently, LB transmits the IDs of the minimum loaded and overloaded VMs to the cloud server. The server then makes a decision regarding the immigration of tasks from the overloaded VM to the less loaded one, focusing on the second hidden layer of deep learning techniques. As a result, the cloud server efficiently balances the workload between VMs with minimal time. This approach proves beneficial in managing a huge number of incoming tasks, leading to minimization of makespan and an increase in throughput.

4. Experimental setup. Experimental evaluation of RITLBDNL and conventional methods, such as TS-DT [1], P2BED-C [2], and RL Approach [3] are implemented using the Java language. To conduct the experiment, we utilize the Personal Cloud Dataset obtained from http://cloudspaces.eu/results/datasets. Major intend of the dataset is to facilitate load balancing. It contains 17 attributes, and 66,245 instances. 17 attributes are row id, account id, file size (task size), operation_time_start, and so on. Two columns, namely time zone and capped, are excluded from the analysis. The aforementioned columns are selected for the purpose of achieving effective load balancing between numerous VMs by big-data CC

5. Performance Analysis. To estimate the performance of RITLBDNL, a comparative analysis was performed between TS-DT [1], P2BED-C [2], and RL Approach [3] in load balancing efficiency, throughput, makespan, response time and memory consumption in Table 1.

Load balancing efficiency: It refers to the ratio of a number of user requests, which are accurately balanced across all VMs. It is computed as given below:

$$LBE = \left[\frac{correctlybalanceduserrequestes}{n}\right] * 100, \quad (13)$$

where LBE indicates a load balancing efficiency, 'n' denotes the total number of user requests.It is measured in percentage (%).

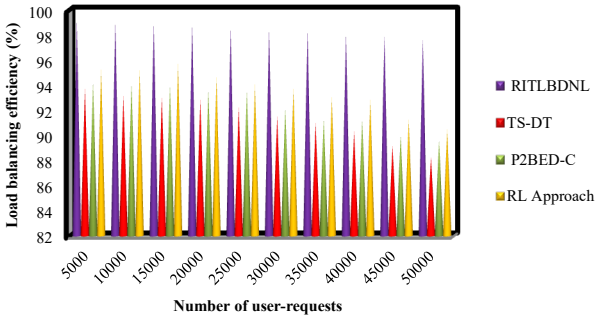


Fig. 5. Analysis of load balancing efficiency

Figure 5 provides a graphical illustration of load balancing efficiency across distinct numbers of user requests ranging from 5,000 to 50,000, taken from the dataset. Figure 4 compares the results of four different algorithms, namely RITLBDNL, TS-DT [1], P2BED-C [2], and RL Approach [3]. It is evident that the RITLBDNL technique yields higher load balancing efficiency. This observation is validated through statistical assessment. In an experiment involving 5000 user requests, the RITLBDNL technique achieved a load balancing efficiency of 99.24%. In contrast, the efficiency of [1], [2], [3] was observed as 93.7%, 94.24%, and 95.2%, respectively. Likewise, different results were attained for every method. Comparing performance outcomes of the proposed method against conventional techniques, overall comparison outcomes show that the RITLBDNL technique increases load balancing efficiency by 8% ,7% and 5% than the [1], [2], [3]. The application of the deep learning technique in RITLBDNL identifies the workload capacity of virtual machines based on resource availability using the Ruzicka Similarity Index function. By utilizing the throttle load-balancing algorithm efficiently, balances workload between VMs, resulting in improved efficiency.

Throughput: it is defined as the ratio of the number of user requests implemented per unit of time in Table 2. It is computed as follows:

$$TP = \left[\frac{Numberofrequestsexecuted}{t\ (seconds)}\right], \tag{14}$$

where 'TP' represents throughput, t indicates time in seconds. It is calculated as requests per second (requests/sec).

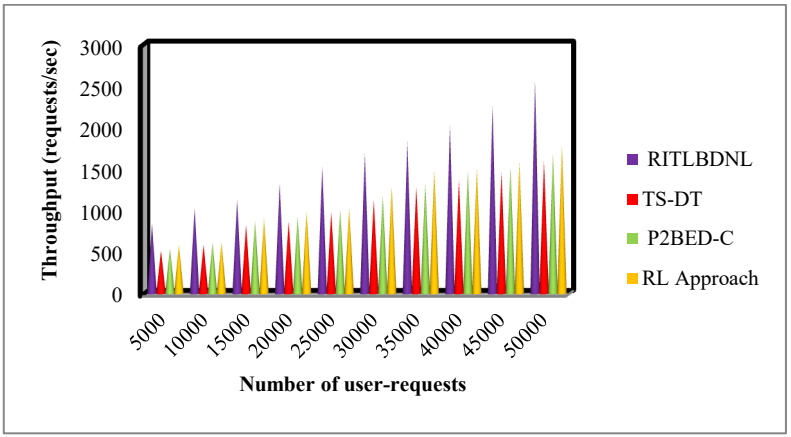


Fig. 6. Analysis of throughput

Figure 6, presented above, illustrates a comparative analysis of throughput. The analysis highlights that the proposed RITLBDNL technique achieved enhanced performance. To ensure the robustness of the RITLBDNL method, ten separate comparisons were conducted for each method. The average of these ten comparisons reveals that the throughput performance using the RITLBDNL technique improved by 54%, 46%, 39% than the [1], [2], [3]. This improvement is achieved through the migration of tasks from the overloaded VMs to the minimum loaded VMs. Consequently, these selected resource-efficient, less loaded virtual machines demonstrate the capability to consistently execute numerous user requests within a specific time.

Makespan: The metric is determined by the duration a virtual machine takes to handle a series of user requests in Table 3. It is calculated as the mathematical dissimilarity among starting as well as completion times of user-requested tasks.

$$M_s = (t_{complete}) - (t_{starting}), \tag{15}$$

where, M_s represents the makespan, $t_{complete}$ indicates request completion time $t_{starting}$de notes a request for finishing time. It is measured in milliseconds (ms).

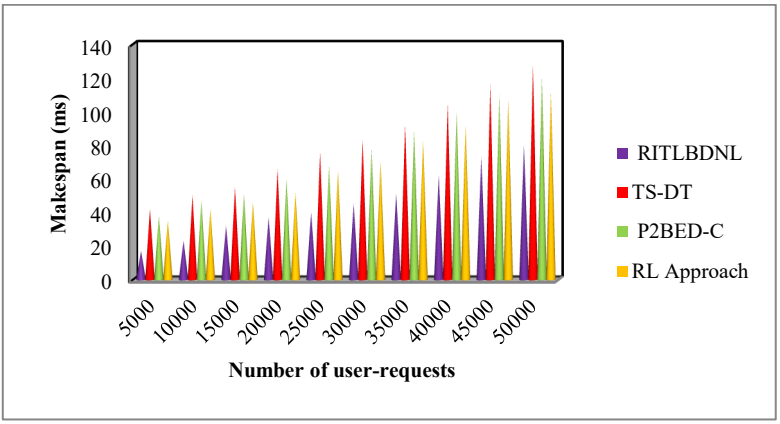


Fig. 7. Analysis of Makespan

Figure 7 depicts a graphical representation of the makespan for load balancing using four methods namely RITLBDNL, TS-DT [1], P2BED-C [2], and RL Approach [3]. The figure illustrates that makespan increases as the number of user requests increases. This occurs as a huge number of user requests during the experiment consumes more time, consequently increasing the makespan. However, in experiments with 5000 user requests, the time taken to complete user requests was only $'17ms'$ using the RITLBDNL technique. The overall makespan was observed to be $42ms$, $38ms$, and $35ms$ for [1], [2], and [3], respectively. Following the experiments, various results were examined for every method. Comprehensive comparison denotes that makespan performance using RITLBDNL is reduced by 45%, 41%, and 36% compared to the existing methods. The RITLBDNL technique employs the Ruzicka similarity index function to analyze the resource status of a VM based on energy, bandwidth, memory, and CPU. Once a minimum loaded VM is identified, LB migrates user requests from an overloaded to a less loaded VM. The less loaded machine requires minimal time to complete the user requests.

Response time: It is defined as the duration it takes to respond user requested tasks in Table 4.

$$RT = n * T\ (\ transmission + waiting + processing), \qquad (16)$$

where RT indicates a response time, n indicates the number of user requests, T represents time for broadcasting, waiting, and processing the user requests. It is measured in milliseconds (ms).

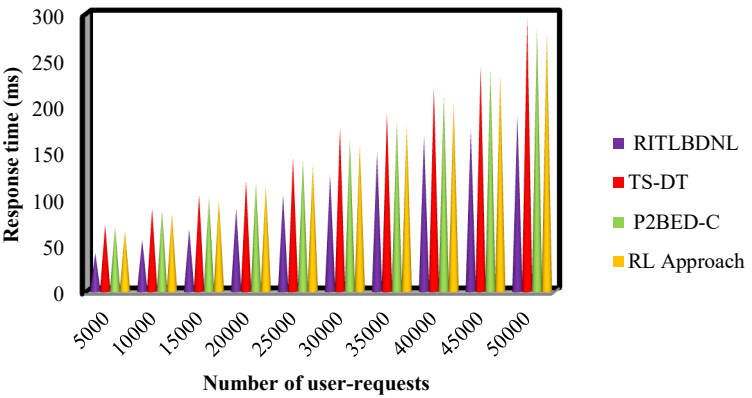


Fig. 8. Analysis of response time

Figure 8 illustrates experimental outcomes of response time with the number of user requests, ranging from 5000 to 50000. As the number of requests enhances, the response time for every method in addition increases. However, the proposed RITLBDNL technique achieves a lower response time compared to existing methods. For instance, with 5000 user requests, the response time for RITLBDNL was observed to be $41.5\,ms$, while [1], [2], and [3] exhibited response times of $72\,ms$, $70\,ms$, and $66\,ms$, respectively. The overall performance results of RITLBDNL are then compared to existing methods, revealing that RITLBDNL minimizes response time consumption by 31%, 28%, and 26% when compared to [1], [2], and [3], respectively. This improvement is achieved using throttled load balancing in the RITLBDNL technique, which effectively performs task migration from overloaded to less loaded virtual machines. Consequently, RITLBDNL minimizes both waiting and processing times for user requests.

5. Discussion. This study compares the proposed RITLBDNL and existing TS-DT [1], P2BED-C [2], and RL Approach [3] based on various parameters, such as load balancing efficiency, throughput, makespan, and response time. The main drawbacks of existing methods such as failure to obtain energy-aware load balancing with a tiny makespan and higher throughput and the failure to employ a resource-aware scheduling approach to assign tasks on VMs. Contrary to existing, Deep Neural Learning and Throttled Load Balancing are utilized in RITLBDNL. By applying this algorithm, the resource status of a virtual machine is examined to find a less

loaded virtual machine. By observing the above table, the results of load balancing efficiency using the proposed RITLBDNL is highly increased than the other existing methods. Also, the response time and makespan of RITLBDNL are greatly reduced than the other works.

Table 1. Comparison table of proposed and existing methods

METHOD	RITLBDNL Technique	TS-DT [1]	P2BED-C [2]	RL Approach [3]
Contribution	To improve load balancing performance using deep neural learning	To allocate tasks using TS-DT	P2BED-C was utilized for data centers	To optimize cloud resource utilization using RL Approach
Merits	Improved throughput and reduced the makespan	Minimized the makespan	Decreased energy consumption	Diminished response time
Demerits	False-positive rate and memory consumption were considered	Energy-aware load balancing was not obtained	Efficiency of the method was not enhanced	Makespan was not efficiently reduced
Load balancing efficiency (%)	98.55	91.5	92.38	93.57
Throughput (requests/sec)	1619.6	1057.9	1113.5	1176.2
Makespan (ms)	46	81	76	70
Response time (ms)	117	165.85	160.7	155.80

Table 1 illustrates a comparison of the proposed RITLBDNL technique and existing TS-DT [1], P2BED-C [2], and RL Approach [3] by using different metrics. Among the three methods, the proposed PCR-AMESCSRO technique provides better performance. The load balancing efficiency was improved by 98.55% using RITLBDNL upon comparison with the three other existing methods. Also, the response time and makespan of the proposed RITLBDNL are obtained as 46 ms and 117 ms which is smaller than the other methods.

6. Conclusion. Balancing the workload is the most important problem in the cloud owing to its dynamic nature. This study introduces a RITLBDNL technique which has been developed to tackle the issue of minimizing makespan and enhancing optimal resource effective load balancing in the cloud. By utilizing the Ruzicka Similarity Index, the

cloud's LB determines the virtual machine resource status for detecting overloaded, less loaded as well as balanced loads. LB performs to dispense workload from heavily loaded virtual machines to minimum loaded ones with higher efficiency. The experimental results also prove that the proposed model has reduced makespan, as well as response time, improved throughput, and efficiency. Compared with various state-of-the-art models, the proposed technique is more efficient. The outcomes of this study have important implications for business applications (i.e., Amazon cloud) to find and classify the resource-efficient VM to allocate the tasks. The less loaded machine needs minimum time and makespan to complete the user requests. Overall, this study provides a valuable contribution to the field of load balancing using DL, and its proposed technique can be extended to other domains where novel DL and optimization are used.

References

1. Mahmoud H., Thabet M., Khafagy M., Omara F. Multiobjective task scheduling in cloud environment using decision tree algorithm. IEEE Access. 2022. vol. 10. pp. 36140–36151.
2. Kumar K. P2BED-C: a novel peer to peer load balancing and energy efficient technique for data-centers over cloud. Wireless Personal Communications. 2022. vol. 123(1). pp. 311–324.
3. Lahande P., Kaveri P., Saini J., Kotecha K., Alfarhood S. Reinforcement Learning approach for optimizing Cloud Resource Utilization with Load Balancing. IEEE Access. 2023. vol. 11. pp. 127567–127577.
4. Nabi S., Ibrahim M., Jimenez J. DRALBA: Dynamic and resource aware load balanced scheduling approach for cloud computing. IEEE Access. 2021. vol. 9. pp. 61283–61297.
5. Sohani M., Jain S. A predictive priority-based dynamic resource provisioning scheme with load balancing in heterogeneous cloud computing. IEEE access. 2021. vol. 9. pp. 62653–62664.
6. Kaviarasan R., Balamurugan G., Kalaiyarasan R. Effective load balancing approach in cloud computing using Inspired Lion Optimization Algorithm. e-Prime-Advances in Electrical Engineering, Electronics and Energy. 2023. vol. 6. DOI: 10.1016/j.prime.2023.100326.
7. Adil M., Nabi S., Aleem M., Diaz V., Lin J. CA-MLBS: content-aware machine learning based load balancing scheduler in the cloud environment. Expert Systems. 2023. vol. 40(4). DOI: 10.1111/exsy.13150.
8. Muthusamy A., Dhanaraj R. Dynamic Q-Learning-Based Optimized Load Balancing Technique in Cloud. Mobile Information Systems. 2023. vol. 2023(1). DOI: 10.1155/2023/7250267.
9. Kruekaew B., Kimpan W. Multi-objective task scheduling optimization for load balancing in cloud computing environment using hybrid artificial bee colony algorithm with reinforcement learning. IEEE Access. 2022. vol. 10. pp. 17803–17818.
10. Mapetu J., Kong L., Chen Z. A dynamic VM consolidation approach based on load balancing using Pearson correlation in cloud computing. The Journal of Supercomputing. 2021. vol. 77(6). pp. 5840–5881.

11. Saroit I., Tarek D. LBCC-Hung: A load balancing protocol for cloud computing based on Hungarian method. Egyptian Informatics Journal. 2023. vol. 24(3). DOI: 10.1016/j.eij.2023.100387.
12. Ashawa M., Douglas O., Osamor J., Jackie R. Retracted Article: Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm. Journal of Cloud Computing. 2022. vol. 11(1). DOI: 10.1186/s13677-022-00362-x.
13. Annie Poornima Princess G., Radhamani A. A hybrid meta-heuristic for optimal load balancing in cloud computing. Journal of grid computing. 2021. vol. 19(2). DOI: 10.1007/s10723-021-09560-4.
14. Mekonnen D., Megersa A., Sharma R., Sharma D. Designing a Component-Based Throttled Load Balancing Algorithm for Cloud Data Centers. Mathematical Problems in Engineering. 2022. vol. 2022(1). DOI: 10.1155/2022/4640443.
15. Haidri R., Alam M., Shahid M., Prakash S., Sajid M. A deadline aware load balancing strategy for cloud computing. Concurrency and Computation: Practice and Experience. 2022. vol. 34(1). DOI: 10.1002/cpe.6496.
16. Pradhan A., Bisoy S., Sain M. Action-Based Load Balancing Technique in Cloud Network Using Actor-Critic-Swarm Optimization. Wireless Communications and Mobile Computing. 2022. vol. 2022(1). DOI: 10.1155/2022/6456242.
17. Udayasankaran P., Thangaraj S. Energy efficient resource utilization and load balancing in virtual machines using prediction algorithms. International Journal of Cognitive Computing in Engineering. 2023. vol. 4. pp. 127–134.
18. Velpula P., Pamula R. EBGO: an optimal load balancing algorithm, a solution for existing tribulation to balance the load efficiently on cloud servers. Multimedia Tools and Applications. 2022. vol. 81(24). pp. 34653–34675.
19. Johora F., Ahmed I., Shajal M., Chowdhory R. A load balancing strategy for reducing data loss risk on cloud using remodified throttled algorithm. International Journal of Electrical and Computer Engineering. 2022. vol. 12(3). pp. 3217–3225. DOI: 10.11591/ijece.v12i3.
20. Khalil M., Shah S., Taj A., Shiraz M., Alamri B., Murawwat S., Hafeez G. Renewable-aware geographical load balancing using option pricing for energy cost minimization in data centers. Processes. 2022. vol. 10(10). DOI: 10.3390/pr10101983.
21. Rajashekar K., Channakrishnaraju Gowda P., Jayachandra A. SCEHO-IPSO: A Nature-Inspired Meta Heuristic Optimization for Task-Scheduling Policy in Cloud Computing. Applied Sciences. 2023. vol. 13(19). DOI: 10.3390/app131910850.
22. Rani P., Singh P., Verma S., Ali N., Shukla P., Alhassan M. An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment. Wireless Communications and Mobile Computing. 2022. vol. 2022. DOI: 10.1155/2022/3365392.
23. Hung L., Wu C., Tsai C., Huang H. Migration-based load balance of virtual machine servers in cloud computing by load prediction using genetic-based methods. IEEE Access. 2021. vol. 9. pp. 49760–49773.
24. Devi K., Sumathi D., Vignesh V., Anilkumar C., Kataraki K., Balakrishnan S. CLOUD load balancing for storing the internet of things using deep load balancer with enhanced security. Measurement: Sensors. 2023. vol. 28. DOI: 10.1016/j.measen.2023.100818.
25. Adaikalaraj J., Chandrasekar C. To improve the performance on disk load balancing in a cloud environment using improved Lion optimization with min-max algorithm. Measurement: Sensors. 2023. vol. 27. DOI: 10.1016/j.measen.2023.100834.

Ellakkiya M. — Research scholar, Department of computer science, Cauvery College For Women (Autonomous); Research scholar, Thanthai Periyar Government Arts and Science

College (Autonomous), Affiliated to Bharathidasan University. Research interests: cloud computing, machine learning. The number of publications — 15. ellakkiya.researchscholar@gmail.com; 620023, Tiruchirappalli, India; office phone: +9994683100.

Ravi T.N. — Associate professor of pg & research, Department of computer science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Affiliated to Bharathidasan University. Research interests: parallel processing, network computing, genetic algorithms, artificial intelligence, data mining. The number of publications — 73. proftnravi@gmail.com; 36/2, Race Course Road, Khajamalai, 620023, Tiruchirappalli, India; office phone: +91(0431)233-1135.

Panneer Arokiaraj S. — Associate professor, Department of computer science, Thanthai Periyar Government Arts and Science College (Autonomous), Affiliated to Bharathidasan University. Research interests: data compression, biometric authentication, data mining. The number of publications — 15. drpancs@gmail.com; 36/2, Race Course Road, Khajamalai, 620023, Tiruchirappalli, India; office phone: +91(0431)242-0079.

УДК 004 **DOI** 10.15622/ia.23.6.9

М. Эллаккия, Т. Рави, С. Паннир Арокиарадж

ИНДЕКСНОЕ РЕГУЛИРУЕМОЕ ГЛУБОКОЕ НЕЙРОННОЕ ОБУЧЕНИЕ РУЖИЧКИ ДЛЯ РЕСУРСОЭФФЕКТИВНОЙ БАЛАНСИРОВКИ НАГРУЗКИ В ОБЛАЧНОЙ СРЕДЕ

Эллаккия М., Рави Т., Паннир Арокиарадж С. **Индексное регулируемое глубокое нейронное обучение Ружички для ресурсоэффективной балансировки нагрузки в облачной среде.**

Аннотация. Облачные вычисления (CC) являются известной технологией, которая позволяет пользователям и организациям получать доступ к сервисам в соответствии с их требованиями. Этот метод вычислений предлагает хранилище, платформы развертывания и подходящий доступ к веб-сервисам через интернет. Балансировка нагрузки является важным фактором оптимизации вычислительных ресурсов и хранения. Она направлена на разумное распределение рабочей нагрузки между каждой виртуальной машиной. Было разработано несколько традиционных методов балансировки нагрузки, которые доступны в литературе. Однако достижение эффективной балансировки нагрузки с минимальным временем завершения и улучшенной пропускной способностью остается сложной задачей. Для повышения эффективности балансировки нагрузки был разработан новый метод, известный как индексированный регулируемый метод Ружички балансировки нагрузки глубокого нейронного обучения (RITLBDNL). Основная цель RITLBDNL состоит в том, чтобы повысить пропускную способность и минимизировать время выполнения работы в облаке. В методе RITLBDNL модель глубокого нейронного анализа включает входной слой, два скрытых слоя и выходной слой для улучшения производительности балансировки нагрузки. На входном слое собираются задачи пользователей облака и отправляются на скрытый слой 1. На этом слое балансировщик нагрузки в облачном сервере анализирует состояние ресурсов виртуальной машины в зависимости от энергии, пропускной способности, объема памяти и ЦПУ с использованием индекса сходства Ружички. Затем виртуальные машины классифицируются как перегруженные, слабо загруженные или сбалансированные. Результаты анализа передаются на скрытый слой 2, где выполняется регулируемая балансировка нагрузки для распределения нагрузки с сильно загруженных виртуальных машин на минимально загруженные. Облачный сервер эффективно распределяет рабочую нагрузку между виртуальными машинами с более высокой пропускной способностью и меньшим временем отклика для обработки огромного количества входящих задач. Для оценки результатов экспериментов предложенный метод сравнивается с другими существующими методами балансировки нагрузки. Результат показывает, что предложенный метод RITLBDNL обеспечивает эффективность балансировки нагрузки с увеличением на 7%, пропускной способностью на 46%, уменьшением времени завершения на 41% и времени отклика на 28% по сравнению с традиционными методами.

Ключевые слова: облачные вычисления, балансировка нагрузки, глубокое обучение, индекс сходства Ружички, регулируемая балансировка нагрузки.

Литература

1. Mahmoud H., Thabet M., Khafagy M., Omara F. Multiobjective task scheduling in cloud environment using decision tree algorithm. IEEE Access. 2022. vol. 10. pp. 36140–36151.

2. Kumar K. P2BED-C: a novel peer to peer load balancing and energy efficient technique for data-centers over cloud. Wireless Personal Communications. 2022. vol. 123(1). pp. 311–324.
3. Lahande P., Kaveri P., Saini J., Kotecha K., Alfarhood S. Reinforcement Learning approach for optimizing Cloud Resource Utilization with Load Balancing. IEEE Access. 2023. vol. 11. pp. 127567–127577.
4. Nabi S., Ibrahim M., Jimenez J. DRALBA: Dynamic and resource aware load balanced scheduling approach for cloud computing. IEEE Access. 2021. vol. 9. pp. 61283–61297.
5. Sohani M., Jain S. A predictive priority-based dynamic resource provisioning scheme with load balancing in heterogeneous cloud computing. IEEE access. 2021. vol. 9. pp. 62653–62664.
6. Kaviarasan R., Balamurugan G., Kalaiyarasan R. Effective load balancing approach in cloud computing using Inspired Lion Optimization Algorithm. e-Prime-Advances in Electrical Engineering, Electronics and Energy. 2023. vol. 6. DOI: 10.1016/j.prime.2023.100326.
7. Adil M., Nabi S., Aleem M., Diaz V., Lin J. CA-MLBS: content-aware machine learning based load balancing scheduler in the cloud environment. Expert Systems. 2023. vol. 40(4). DOI: 10.1111/exsy.13150.
8. Muthusamy A., Dhanaraj R. Dynamic Q-Learning-Based Optimized Load Balancing Technique in Cloud. Mobile Information Systems. 2023. vol. 2023(1). DOI: 10.1155/2023/7250267.
9. Kruekaew B., Kimpan W. Multi-objective task scheduling optimization for load balancing in cloud computing environment using hybrid artificial bee colony algorithm with reinforcement learning. IEEE Access. 2022. vol. 10. pp. 17803–17818.
10. Mapetu J., Kong L., Chen Z. A dynamic VM consolidation approach based on load balancing using Pearson correlation in cloud computing. The Journal of Supercomputing. 2021. vol. 77(6). pp. 5840–5881.
11. Saroit I., Tarek D. LBCC-Hung: A load balancing protocol for cloud computing based on Hungarian method. Egyptian Informatics Journal. 2023. vol. 24(3). DOI: 10.1016/j.eij.2023.100387.
12. Ashawa M., Douglas O., Osamor J., Jackie R. Retracted Article: Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm. Journal of Cloud Computing. 2022. vol. 11(1). DOI: 10.1186/s13677-022-00362-x.
13. Annie Poornima Princess G., Radhamani A. A hybrid meta-heuristic for optimal load balancing in cloud computing. Journal of grid computing. 2021. vol. 19(2). DOI: 10.1007/s10723-021-09560-4.
14. Mekonnen D., Megersa A., Sharma R., Sharma D. Designing a Component-Based Throttled Load Balancing Algorithm for Cloud Data Centers. Mathematical Problems in Engineering. 2022. vol. 2022(1). DOI: 10.1155/2022/4640443.
15. Haidri R., Alam M., Shahid M., Prakash S., Sajid M. A deadline aware load balancing strategy for cloud computing. Concurrency and Computation: Practice and Experience. 2022. vol. 34(1). DOI: 10.1002/cpe.6496.
16. Pradhan A., Bisoy S., Sain M. Action-Based Load Balancing Technique in Cloud Network Using Actor-Critic-Swarm Optimization. Wireless Communications and Mobile Computing. 2022. vol. 2022(1). DOI: 10.1155/2022/6456242.
17. Udayasankaran P., Thangaraj S. Energy efficient resource utilization and load balancing in virtual machines using prediction algorithms. International Journal of Cognitive Computing in Engineering. 2023. vol. 4. pp. 127–134.

18. Velpula P., Pamula R. EBGO: an optimal load balancing algorithm, a solution for existing tribulation to balance the load efficiently on cloud servers. Multimedia Tools and Applications. 2022. vol. 81(24). pp. 34653–34675.
19. Johora F., Ahmed I., Shajal M., Chowdhory R. A load balancing strategy for reducing data loss risk on cloud using remodified throttled algorithm. International Journal of Electrical and Computer Engineering. 2022. vol. 12(3). pp. 3217–3225. DOI: 10.11591/ijece.v12i3.
20. Khalil M., Shah S., Taj A., Shiraz M., Alamri B., Murawwat S., Hafeez G. Renewable-aware geographical load balancing using option pricing for energy cost minimization in data centers. Processes. 2022. vol. 10(10). DOI: 10.3390/pr10101983.
21. Rajashekar K., Channakrishnaraju Gowda P., Jayachandra A. SCEHO-IPSO: A Nature-Inspired Meta Heuristic Optimization for Task-Scheduling Policy in Cloud Computing. Applied Sciences. 2023. vol. 13(19). DOI: 10.3390/app131910850.
22. Rani P., Singh P., Verma S., Ali N., Shukla P., Alhassan M. An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment. Wireless Communications and Mobile Computing. 2022. vol. 2022. DOI: 10.1155/2022/3365392.
23. Hung L., Wu C., Tsai C., Huang H. Migration-based load balance of virtual machine servers in cloud computing by load prediction using genetic-based methods. IEEE Access. 2021. vol. 9. pp. 49760–49773.
24. Devi K., Sumathi D., Vignesh V., Anilkumar C., Kataraki K., Balakrishnan S. CLOUD load balancing for storing the internet of things using deep load balancer with enhanced security. Measurement: Sensors. 2023. vol. 28. DOI: 10.1016/j.measen.2023.100818.
25. Adaikalaraj J., Chandrasekar C. To improve the performance on disk load balancing in a cloud environment using improved Lion optimization with min-max algorithm. Measurement: Sensors. 2023. vol. 27. DOI: 10.1016/j.measen.2023.100834.

Эллаккия М. — научный сотрудник, факультет компьютерных наук, Колледж Кавери для женщин (автономный); научный сотрудник, Государственный колледж искусств и наук имени Тхантхай Перияра (автономный), филиал Университета Бхаратхидасан. Область научных интересов: облачные вычисления, машинное обучение. Число научных публикаций — 15. ellakkiya.researchscholar@gmail.com; 620023, Тируччираппалли, Индия; р.т.: +9994683100.

Рави Т.Н. — доцент кафедры, факультет компьютерных наук, Колледж Джамаля Мохаммеда (автономный), Тируччираппалли, филиал Университета Бхаратхидасан. Область научных интересов: параллельная обработка, сетевые вычисления, генетические алгоритмы, искусственный интеллект и интеллектуальный анализ данных. Число научных публикаций — 73. proftnravi@gmail.com; Ипподром, Хаджамалай, 36/2, 620023, Тируччираппалли, Индия; р.т.: +91(0431)233-1135.

Паннир Арокиарадж С. — доцент кафедры, кафедра компьютерных наук, Государственный колледж искусств и науки Тантай Перияр (автономный), Университет Бхаратидасан. Область научных интересов: сжатие данных, биометрическая аутентификация, интеллектуальный анализ данных. Число научных публикаций — 15. drpancs@gmail.com; Ипподром, Хаджамалай, 36/2, 620023, Тируччираппалли, Индия; р.т.: +91(0431)242-0079.

DOI 10.15622/ia.23.6.10

C. NARAYANARAO, V. MANDAPATI, B. BODDU

SYNERGISTIC APPROACHES TO ENHANCE IOT INTRUSION DETECTION: BALANCING FEATURES THROUGH COMBINED LEARNING

Narayanarao C., Mandapati V., Boddu B. **Synergistic Approaches to Enhance IoT Intrusion Detection: Balancing Features through Combined Learning.**

Abstract. The Internet of Things (IoT) plays a crucial role in ensuring security by preventing unauthorized access, malware infections, and malicious activities. IoT monitors network traffic as well as device behaviour to identify potential threats and take appropriate mitigation measures. However, there is a need for an IoT Intrusion Detection system with enhanced generalization capabilities, leveraging deep learning and advanced anomaly detection techniques. This study presents an innovative approach to IoT IDS that combines SMOTE-Tomek link and BTLBO, CNN with XGB classifier which aims to address data imbalances, improve model performance, reduce misclassifications, and improve overall dataset quality. The proposed IoT IDS system, using the IoT-23 dataset, achieves 99.90% accuracy and a low error rate, all while requiring significantly less execution time. This work represents a significant step forward in IoT security, offering a robust and efficient IDS solution tailored to the changing challenges of the interconnected world.

Keywords: min-max normalization, SMOTE-Tomek Link, BTLBO algorithm, CNN with XGB, Adam Optimizer.

1. Introduction.

IoT is a system that connects items that may transfer information without the necessity for interactions between humans and computers, such as computing equipment, automated and digital technologies, objects, animals, and humans. IoT essentially connects the real and virtual worlds. IoT's primary idea is to create a secure, independent link that allows info interchange between actual physical devices and applications [1]. The IoT Analytics study reveals that over 11 billion IoT devices are connected and utilized. Additionally, it is demonstrated the number of devices has increased by more than 10%. It is predicted that over 21 billion linked IoT devices will be worldwide by 2025. Due to general use in several areas and businesses, such as agriculture, transportation, logistics, and healthcare are all examples of smart cities and smart homes, the IoT has seen tremendous growth in recent years [2]. Organisations that use IoT devices in information technology systems have introduced new cybersecurity risks. These new threats call into question fundamental assumptions such as operating ecosystem security, mobility, efficiency, and safety. New danger vectors risk financial and bodily well-being and affect lives' technical components [3, 4].

The ecosystem in which IoT devices are deployed is vulnerable to several threats [5] from outsiders, including hackers, malicious software, and viruses [6]. These hackers' primary objective is to launch assaults

compromising network data integrity [7]. Additionally, the intrusion may result in a denial of service (DoS) attack [8] that depletes energy in an IoT environment as well as network and device resources [9]. The author infers from the literature that many studies on the IoT employ security methods based on cryptography, such as symmetric and public key cryptosystems [10, 11].

Because IoT devices have limited resources, implementing cryptographic algorithms in IoT security leads to effective communication and processing overhead [12]. This problem can be resolved by designing and implementing intrusion detection systems (IDS). To effectively secure IoT communication, IDS has been accepted in IoT environments to guide and detect imposters [13, 14]. The detection of critical and particular threats for traditional networks and a portion of the Internet of Things networks has been taught using various Machine learning (ML) and Deep learning (DL) models for IDS [15]. IDSs currently apply similar attribute ideas to IoT devices. However, IoT devices vary in a variety of bearings, including physical characteristics, utility, potential computing power, and variable capacity aimed at generating decided appearances [16, 17]. When hubs are merged then generate data, the features become sparse as unimportant qualities are set to null values or zeros. One limitation striking the precision's effectiveness is data sparsity. A selection of features, an essential component of an ML method, contributes much to the training phase speed and finding accuracy. To enhance the identification of variations of anomalous behavior, many feature selection strategies have been developed. However, the accuracy of anomaly-based ID detection is considered a significant issue due to the constantly evolving nature of the IoT ecosystem. To achieve robust performance across the varied IoT environment, this study provides a unique technique for deep learning IDS.

The primary contribution of the proposed project is given below.

– IoT faces significant security challenges due to remote access and unreliable networks. To prevent attacks, IoT environments employ effective security management techniques and ID systems. Still, there are also possibilities to improve both accuracy and performance. The proposed novel deep learning technique in the IDS aims to address this issue.

– For the preprocessing stage, the novel approach utilizes the process of removal of null values, one hot encoding technique incorporating the numerical representation which enables the creation of a digital feature vector, Minmax normalization for dimensional removal, and Synthetic Minority Over-sampling Technique (SMOTE) Tomek technique for balancing synthetic data.

– This study employs a novel technique called Binary Teaching Learning Based Optimization (BTLBO) algorithm for feature selection.

– For feature extraction and classification, the approach uses a network as Convolution Neural Network (CNN) with extreme Gradient Boosting (XGB) classifier to predict the classes. Thus, the hybrid method accurately detects the intrusion in the IoT.

The description of the sections indicates that the article will cover existing research on IoT intrusions, the proposed methodology and results of the proposed work, and also draw conclusions and suggest some possible further research directions.

2. Literature survey. This portion of the article delves into CNN-based intrusion detection systems published in the literature.

To resolve network assault binary and multiclass categorization, in paper [18] the authors created two models based on DL and employed a CNN architecture. A hybrid two-step pre-processing method is also suggested to produce useful features. Deep feature synthesis is used in the proposed strategy to combine dimensionality reduction with feature engineering. It was shown that the multiclass models' accuracy of classification is lower than binary class models. Instead of traditional anomalous attack behaviors, the authors in paper [19] used statistical behaviors since they are simpler to reckon and extract without sacrificing performance. Because it primarily considers statistical characteristics of network traffic, the model's accuracy for multiclass categorization is lower.

The Temporal CNN (TCNN) was proposed by the authors in paper [20] and is accumulated with SMOTE with nominal continuity to handle imbalanced datasets. To find network system abnormality, the authors in [21] used the novel CNNs binary and multiclass classification model. Even though CNN has various characteristics that make it especially suited for IDS, such as high attainment precision, finding rate, model training time, and feature selection procedures, the efficiency of ML models is improved. Hybrid models have been quite common in recent years for categorizing attacks on IoT networks. To address the IDS issues related to time consumption and inefficiency, the authors in study [22] introduced a cascade ID method that depends on distributed k-means and Random Forest.

Along with the Ant Lion optimization approach, which combines CNN and Long Short Term Memory (LSTM), study [23] introduced a new customized recurrent neural network model that is optimized for detecting intrusion. The Lion Swarm Optimization method is employed to optimize CNN hyperparameters for perfect composition for learning structural data. The authors in [24] suggested a highly accurate IDS model for valuable

uprooting and learning of contiguous secular features using optimized CNN and Hierarchical multiscale LSTM. Through careful feature selection, this approach can increase detection accuracy. A Deep Capsule Network (DCN) ID model that depends on the system of attention was suggested by the authors in [25]. To increase feature extraction, the model integrates DCN, and an Attention Mechanism is employed to minder the model's attention toward qualities with substantial consequences. Two solutions are utilized to balance the dynamic powerful routing procedure after the double routing algorithm captures the characteristics in multiple directions. Because the dynamic routing contrivance of the CN consumes more time than a normal NN, the operational efficiency of CN must be increased.

To protect the computer, network nodes and data, in study [26] the authors introduced a unique Network Intrusion Detection System architecture that depends on a deep capsule neural network that creates usage of network spectrogram pictures produced utilizing the short-time Fourier transform. In comparison to previous published works, the computational complexity is higher. To identify intrusions in the IoT environment, in paper [27] the authors found a novel multi-objective evolutionary CNN for IDS. In the context of IoT and cloud computing, a new approach to IDSs was proposed in [28]. The major goal is to develop effective feature extraction and selection strategies by utilizing the widespread use of deep learning and metaheuristic optimization algorithms. An approach based on PCA and CNN was put out by the authors in paper [29] to identify intrusion in EDGE Computing. Feature selection and data balancing are not employed to improve categorization accuracy. For machine learning-based IDS, the authors in [30] gave a feature selection technique for extracting useful subsets of features based on the idea of the math concept of sets. The designed ML-based IDS system contains three stages: data pre-processing, proportions lessening and size selection, model training, and categorization.

An analysis of various works reveals that many of them neglect the multiclass imbalance distribution and feature selection techniques for improved accuracy. Unique techniques are needed to manage imbalance distribution and select the best feature set for multiclass classification.

3. Proposed methodology. IoT IDS face limitations due to data imbalances, model generalization, and performance optimization. Traditional approaches struggle to address these issues, leading to suboptimal performance and limited scalability. Imbalanced datasets can bias model training and result in poor classification performance, especially for minority intrusion classes. Current techniques for handling imbalanced data may not capture underlying patterns or introduce biases. Conventional

IDS methods also lack the ability to generalize across diverse IoT environments and adapt to evolving threats, relying on handcrafted features or simplistic anomaly detection algorithms. To overcome these limitations, our research proposes an efficient ID using a deep learning-based categorization strategy to increase the IDS's accuracy. The proposed method starts with pre-processing to eliminate duplicate instances and missing values, followed by numerical processing to produce a digital feature vector. The Min-Max Normalization approach is used for linear and uniform mapping of feature ranges, characteristics can be adjusted for faster removal of dimensions and arithmetic processing.

The problem with ML-based IDSs is using an unbalanced dataset to train a model. SMOTE-Tomek links, which combine SMOTE for artificial information for the minority class, along with Tomek Connections for excluding data identified as from the majority, solve this issue. The proposed study uses CNN with XGB for classification, which includes two convolutional hidden layers, batch normalization, Exponential Linear Unit (ELU), max-pooling, dropout layer, and Adam optimizer weights. Our work aims to enhance the performances of IoT IDs. Figure 1 depicts the overall design of the proposed technique.

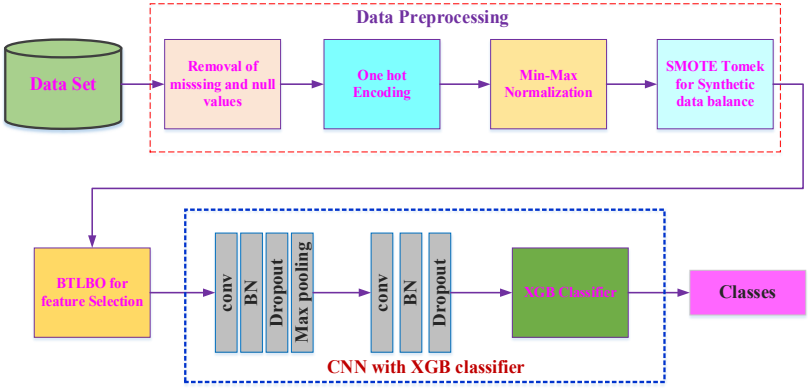


Fig. 1. Proposed Block Diagram

3.1. Data Preprocessing. It is a crucial stage in data analysis and ML work, involving cleaning, conversion, and converting raw information into suitable formats for analysis or model training. Data preparation can have a substantial impact on the accuracy and effectiveness of an analysis or model. It involves removing missing and null values from the dataset to ensure the effectiveness of the ML model, as this removes incomplete or unreliable data.

Informal information, including categorical and symbolic aspects, is handled using a one-hot encoding technique. This process converts non-numeric data into a digital feature vector, which can interpret and utilize these features effectively.

The Min-Max Normalization method is used for smooth and consistent mapping of each feature's parameter variation over a specific interval. This normalization technique improves the stability and convergence of machine learning algorithms by ensuring consistent scaling across features. It is a way that provides a balance of assessments among information obtained from prior and subsequent procedures. Features are further normalized to a Gaussian distribution, which aids in the removal of dimensions and accelerates arithmetic processing.

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)}, \quad (1)$$

where $X_{new} = the\ new\ value\ from\ the\ normalized\ results$, X = old value, $Max(X) = Maximum\ value\ in\ the\ dataset$, $Min(X) = Minimum\ value\ in\ the\ dataset$.

3.1.1. Proposed SMOTE-Tomek Link. The research on strengthening IoT IDS through synergistic techniques is focused on overcoming the issues given by data imbalances and improving the overall performance of detection models. One of the innovations in this study is the use of an upgraded SMOTE-Tomek link methodology that outperforms existing methods. The SMOTE technique, when paired with Tomek links, helps to balance the dataset by producing synthetic samples for the minority classes and deleting overlapping samples between the classes. This method is unique because it balances the data, guaranteeing that the created synthetic samples contribute positively to the model's learning process.

This technique is a new approach to ML-based IDSs, which addresses the issue of unbalanced datasets, ensuring accurate identification of the minority class instances which combines the SMOTE to generate synthetic data for the minority class and Tomek Links [46] to remove Tomek links from the majority class, resulting in more accurate model training. This strategic combination of preprocessing techniques contributes to the overall efficacy and reliability of the IDS model, making it more suitable for real-world scenarios with class imbalances. The use of this technique is critical to overcoming the challenge of an imbalanced dataset. The proposed technique addresses class imbalance by combining SMOTE for synthetic data generation and Tomek Links for strategic data removal. This results in more robust and accurate IDS model training.

SMOTE-Tomek combines these two techniques in the following way.

1. Start with the original dataset, which may be imbalanced.
2. Apply SMOTE to oversample the minority class, creating synthetic instances for it.
3. After SMOTE, the dataset may still contain the Tomek link, it denotes the proximity of two instances from different classes.
4. Locate and eliminate any Tomek linkages from a dataset.

The SMOTE-Tomek link aims to enhance dataset balance by oversampling the minority classes and removing noisy samples, enhancing classification model performance. It is particularly effective in imbalanced datasets and can improve machine learning models. However, it may not be suitable for all classification problems. The used resampling technique is determined by the dataset's properties and the machine learning task objectives. The effectiveness of SMOTE-Tomek should be evaluated through cross-validation and relevant performance metrics. After balancing the data with class-wise sampling, it should be proceeded to the process of feature selection by using BTLBO.

3.2. BTLBO algorithm. It is a powerful technique for feature selection in machine learning and data analysis. It optimizes the subset of relevant features used for training models, focusing on refining model performance and interpretability. Drawing inspiration from the teaching-learning process, BTLBO uses a binary encoding scheme to efficiently explore the solution space, evaluating, and selecting features that significantly contribute to the model's predictive power. BTLBO is a novel and efficient approach to feature selection, overcoming limitations in traditional methods such as high-dimensional datasets, computational inefficiency, and suboptimal search strategies. Its binary representation and amalgamation of teaching and learning strategies enable systematic evaluation of feature subsets, enhancing model accuracy and interpretability in ML applications. This innovative method offers a robust solution for addressing the challenges faced by traditional methods.

Let M_i be the mean, and T_i be the teacher at any iteration i. T_i will try to move mean M_i towards its own level. First students are trained with the assistance of a teacher. Assume that ‘s’ represents the number of features in each iteration k. Attributes, {f=1,2,….s}, ‘t’ is the number of instances i.e., population, individuals, {i=1,2,…t}.

BTLBO is a teaching-learning optimization algorithm that stimulates the teaching and learning process in a classroom. It involves all students as a population, with subjects similar to decision variables. The best learner is considered a teacher, who transfers knowledge to all learners. The learner's

performance is based on the fitness value of the individual in the population. TLBO operates in two phases: teacher and learner (Figure 2).

Algorithm 1. BTLBO [42]

Step 1: Initialize several instances (binary population), several subscripts and $X_{f,i,k}$ and an end condition.
Step 2: Calculate the mean for learners as $M_{f,k}$
Step 3: Using equation (2), determine the fitness of people.

$$\text{Fitness } (X_{f,i,k}) = \text{Accuracy } (X_{f,i,k}). \quad (2)$$

(Teacher Phase)

Step 4: Upgrade students with the assistance of the instructor. i.e. teacher phase

(a) Choose the highest fitness value from the group as a teacher.

(b) Calculate the mean variation for all traits concerning the best individual as shown in the equation

$$\text{Diff\_}Mean_{f,i,k} = r_k(X_{f,i,best,k} - T_f M_{f,k}), \quad (3)$$

where $X_{f,i,best,k}$ the best individual in f. T_f, teaching factor with the value 1 or 2, r_k is the random number ranging from 0 to 1.

(c) The best person serves as a teacher and mentor to others.

(d) $X'_{f,i,k} = 0, if\ X_{f,i,k} + \text{Diff\_}Mean_{f,k} < 0.5$

$$X'_{f,i,k} = 1, if\ X_{f,i,k} + \text{Diff\_}Mean_{f,k} \geq 0.5, \quad (4)$$

where $X'_{f,i,k}$ the trained value of $X_{f,i,k}$

If the result $X'_{f,i,k}$ is better than $X_{f,i,k}$, Otherwise, replace the previous value with the new value.

Step5: updates each learner with the assistance of other learners using the eq (5,6)

(Learner Phase)

(a) Select two cases U and V that satisfy the criterion $X'_{total-U,k} \neq X'_{total-V,k}$ at random, Where X_(total-U,k)^' , X_(total-V,k)^' of U and V respectively

(b) If $X'_{total-U,k}$ is better than $X'_{total-V,k}$

$$X''_{f,U,k} = 1 \text{ if } X'_{f,U,k} + r_k(X'_{f,U,k} - X'_{f,V,k}) \geq 0.5). \quad (5)$$

Or

$$X''_{f,U,k} = 0 \text{ if } X'_{f,U,k} + r_k + (X'_{f,V,k} - X'_{f,U,k}) < 0.5.$$
$$X''_{f,U,k} = 1 \text{ if } X'_{f,U,k} + r_k + (X'_{f,V,k} - X'_{f,U,k}) \geq 0.5. \quad (6)$$

(c) If X_(f,U,k)^"= is better than X_(f,U,k)^', then continue the prior value, otherwise, substitute the previous value.

Step 6: if the stop condition is pleased, then report the result, then go the second step

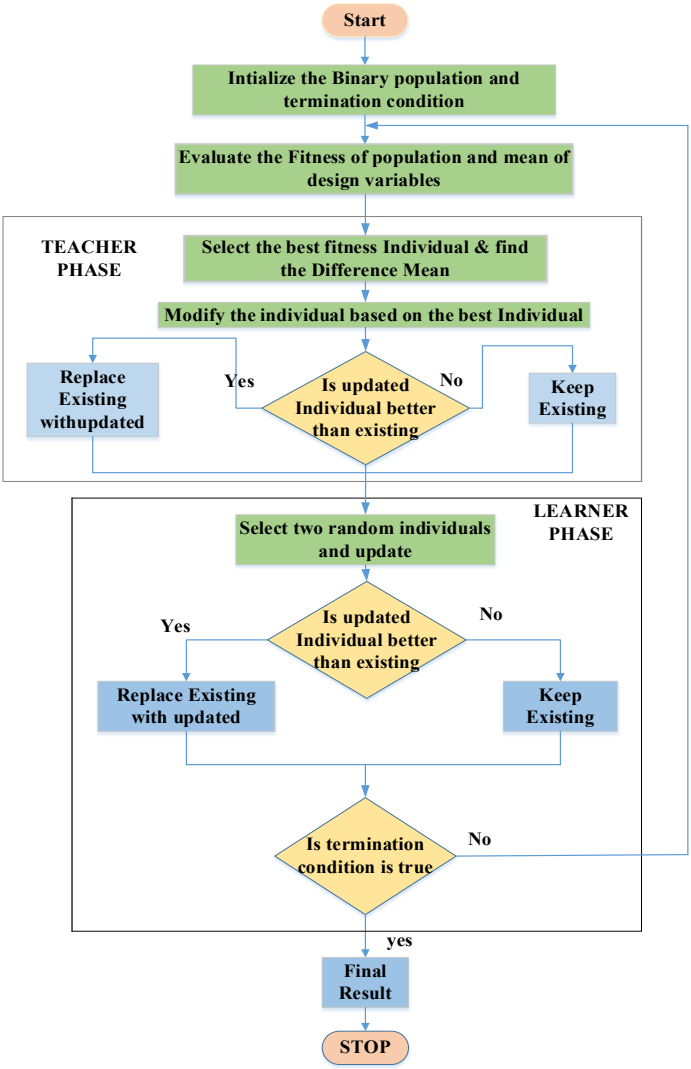


Fig. 2. BTLBO Algorithm Flow Chart [40]

This approach uses binary bits 1 and 0 to signify the presence or absence of a characteristic in a population, with the length of the binary string corresponding to the number of levels in each dataset. In the instructor phase, the mean value reflects the likelihood of a feature's appearance, while the difference mean describes learners' variation. The

teacher with the highest accuracy is named. Wrapper-based feature selection approaches use predictive models to evaluate population fitness, with classification accuracy as a fitness value. The teaching factor will be chosen at random from 1 to 2. The classification accuracy can be expressed as

$$CA = \text{correct classified Instances}/\text{Total Instances}. \quad (2)$$

An individual with the lowest error rate or the highest accuracy will determine the last solution of optimal characteristics selection. The individuals added to the dataset are utilized to teach NN to boost effectiveness. The process for the optimum number of iterations is executed, and the method flow as shown in Figure 2 is presented. The BTLBO algorithm was created in the study to select an optimal subset of characteristics from an extensive database. Once features are selected, they are then refined into a categorization pipeline using CNN and the XGB classifier, combining the strengths of both techniques to improve classification performance.

3.3. Hybrid CNN-XGB Classifier. The proposed method combines CNN with the XGB classifier, resulting in a powerful combination for effective classification tasks. This hybrid approach is chosen to take advantage of CNN's capabilities in capturing intricate spatial hierarchies and patterns from complex data, such as images or sequences. CNN excels at automatic feature extraction, producing high-level representations required for reliable classification [44]. The subsequent use of the XGB classifier adds a layer of ensemble learning, allowing for efficient handling of nonlinear relationships and improving overall model performance. The proposed method enhances classification accuracy and interpretability by using CNN for feature extraction and XGB for ensemble-based classification. This approach addresses limitations in previous research, such as high-dimensional data handling and complex feature relationships, by combining deep learning and gradient boosting for optimal predictive accuracy. The CNN structure is in Figure 3.

The input layer is followed by another convolutional and pooling layer, which is a sub-sampling layer. The pooling layer then pools relevant features and performs the extraction function, and the unused features are clarified out in convolutional and pooling layers. This work presents a new intelligent deep classification algorithm using the CNN algorithm with IF…THEN rules. [45] The CNN XG classifier conducts fusion and maximal pooling operations, representing convolution for a pair of functions f, g using an integral equation for the operator t.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau. \quad (3)$$

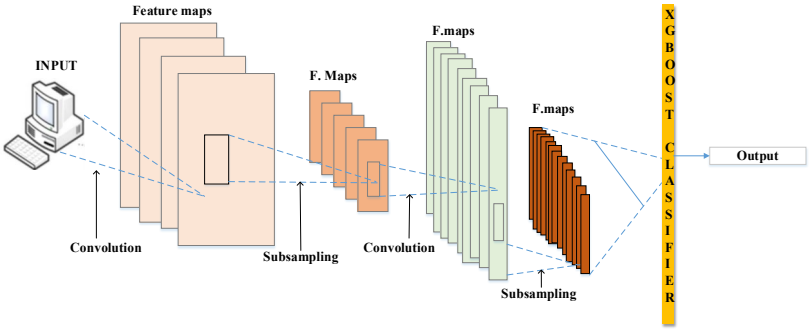


Fig. 3. Representation of the combined proposed technique

The CNN consists of dual convolutions, two batch normalizations, two dropouts and one max pooling layer, all of which are designed to automatically capture hierarchical representations from input data. Once the CNN has extracted these complex features, the feature set is fed into the XGB classifier for the final classification stage.

$$f(x) = \left[\frac{(x+1)}{x}\right]. \quad (4)$$

The XGB classifier excels at leveraging ensemble learning via gradient boosting, which creates a series of decision trees that collectively improve the model's predictive capabilities. The features extracted by the CNN serve as high-level representations of IoT data, capturing patterns and complexities that are critical for distinguishing between normal and anomalous network behaviour. The XGB classifier, which is adept at handling complex relationships and non-linearities, refines these features using a boosting process. The XGB algorithm uses CNN-extracted features to refine decision boundaries in a classification workflow. Each decision tree in the ensemble contributes to the overall classification decision, combining the strengths of multiple weak learners. This integrated approach ensures the model learns intricate features and refines them through the ensemble-based learning strategy of XGB. The final classification output distinguishes between normal and malicious IoT activities, providing a reliable and adaptive intrusion detection system for IoT environments.

4. Results and discussion. The study highlights the effectiveness of the synergistic approach to IoT Intrusion Detection System, which integrates

SMOTE-Tomek link, BTLO, Convolutional Neural Network, and XGB classifier. This section provides the dataset description, performance of the proposed method, evaluation metrics, and comparison. During the process, a Python tool is used for implementation with the tensor flow library.

4.1. Dataset description. The IoT-23 [41], [43] dataset is a crucial resource for IoT security and intrusion detection research, derived from real-world devices, simulated environments, and network traffic captures. It provides comprehensive insights into IoT operations, including network traffic attributes, device-specific information, and normal behavior. The dataset is essential for model training and evaluation, as instances are labelled to indicate potential intrusions. The analysis is carried out using IoT-23 dataset consists of 20 malware catches accomplished in IoT devices and 3 captures for benign IoT device traffic. The IoT-23 dataset consists of twenty-three different IoT network traffic recordings called scenarios. These scenarios are divided into 20 network captures of pcap files from infected IoT devices in the name of the malware executed on each scenario and three network captures of real IoT device network traffic. The dataset, which covers a wide range of anomalies in IoT ecosystems, is crucial for robust intrusion detection models due to its potential for class imbalance, a common challenge in real-world datasets.

4.2. Experimental Settings. For the evaluation of this research, the total number of samples is (1211513, 31). After preprocessing by removing the zeros and null value, the number of samples is (981934, 31). The proposed technique split the samples into 75% (736450, 31) for training and 25% (245484, 31) for testing. The number of class instances was used to calculate class weights, so the class with the fewest instances will have a high weight. Each CNN model was trained with a batch size of 32 and 10 iterations using the Adam optimizer with a learning rate of 0.001 for 100 epochs. Early halting reduces the possibility of excessive fitting, which happens when a model is refined over an abundance of eras. The batch size increased and a number of epochs lowered to see if the model's accuracy improved. For training and validation sets, the precision and loss of each model were evaluated at each epoch value.

4.3. Metrics for evaluation. The performance metrics are evaluated using the following values:

Accuracy:

$$Accuracy = \frac{TP + TN}{number\ of\ all\ samples\ in\ the\ testing\ sets}. \quad (5)$$

Precision:

$$Precision = \frac{TP}{TP+FP}. \quad (6)$$

Recall:

$$Recall = \frac{TP}{TP + FN}. \quad (7)$$

The value recall is equal to the sensitivity value.

Specificity: relates to how successfully a classifier can identify bad outcomes.

$$Specificity = \frac{TN}{TN + FP}. \quad (8)$$

The value of specificity is equal to the True Negative Rate (TNR) value.

F 1:

$$Fmeasure = \frac{2 \times precision \times recall}{precision+recall} = \frac{2TP}{2TP+FP+FN}. \quad (9)$$

***PPV*:** termed a positive predictive value, which is calculated by

$$PPV = \frac{TP}{TP + FP}. \quad (10)$$

***NPV*:** negative predictive value, which is calculated by

$$NPV = \frac{TN}{TN + FN}, \quad (11)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative.

4.4. Evaluation Findings and Comparisons. To detect intrusions, the experiment used the CNN_XGG algorithm. Sensitivity, Specificity, PPV, and NPV results for multiTable.1 provides the experimental result of the proposed work which was compared with three different learning-based IDS models that work in the IoT-23. Each subset is used to evaluate CNN_LSTM, CNN_BiLSTM, CNN_GRU, and also proposed CNN_XGB models. The accuracy, precision, recall, and F1, Sensitivity, Specificity, PPV, and NPV score, of the IoT-23 dataset using CNN_XGB by comparing

with the existing ranking algorithm [37] of these models are presented in Figure 6. CNN_XGB model performs better than existing models. A single hidden layer CNN successfully classified normal and abnormal situations in the IoT-23 dataset [37], demonstrating its ability to learn meaningful patterns from network traffic data, making this result impressive in detecting normal and anomalous occurrences in the IoT-23 dataset.

In the training phase, the weights for classes were computed according to the number of occurrences for every group; the minority class with a small number of instances will receive better priority. SMOTE followed to correct the class imbalances. The evaluation of the proposed model is shown in Table 1.

Table 1. Performance assessment of the proposed model

Metrics	Proposed model
Accuracy	99.90 %
Precision	99.51 %
Recall	99.95 %
F1	99.20 %
Sensitivity	99.91 %
Specificity	100 %
PPV	99.92 %
NPV	100 %
Error rate	0.012

The novel model was validated by separating the dataset presenting accuracy performances of CNN and XGB algorithms in Figure 4, the training loss of the DL algorithm mentions a link between training loss and the number of epochs in the proposed work (Figure 5). The comparative evaluation of IoT Intrusion Detection Systems is shown in Table 2.

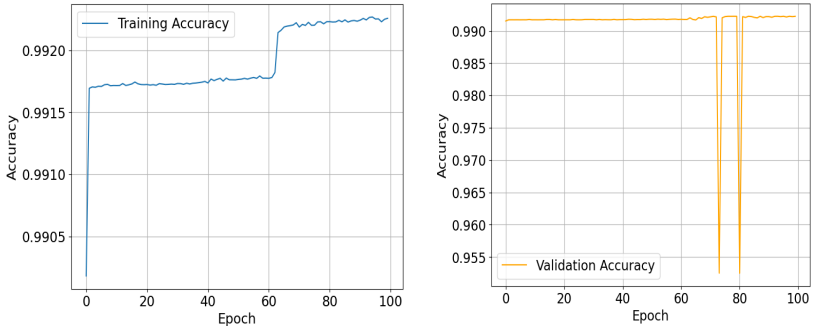


Fig. 4. Training/Validation accuracy

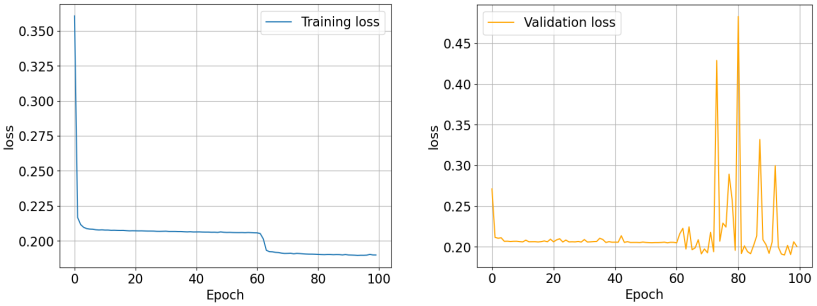


Fig. 5. Training Vs validation loss

Table 2. Performance Comparison of IoT IDS Approaches of multi-class classification

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
CNN_LSTM[37]	99.83	99.11	98.92	99.01	99.83	99.96	99.83	99.98
CNN_BiLSTM[37]	87.99	99.29	97.87	98.56	98.87	99.97	99.87	99.99
CNN_GRU[37]	86.99	99.18	99.01	99.09	99.86	99.98	99.86	99.98
Proposed	99.90	99.51	99.95	99.20	99.91	100	99.92	100

Table 2 shows that the proposed DL-based IDS models obtained better performance in identifying various forms of cyberattacks compared to existing features. In research paper [38], the RNN model gained an accuracy of 98.31%, so it was concluded that a DL model might considerably improve accuracy, permitting efficient security against threats in IoT systems. The suggested CNN model's accuracy was very low. However, the proposed model CNN_XGB showed high accuracy compared to other multi-class classifiers. From Table 3, it can be seen that the accuracy rates of DL-based IDS models are comparable, which shows the proposed model achieves the lowest error rate among IDS models [39] belonging to the CNN_XG ensemble by 0.012.

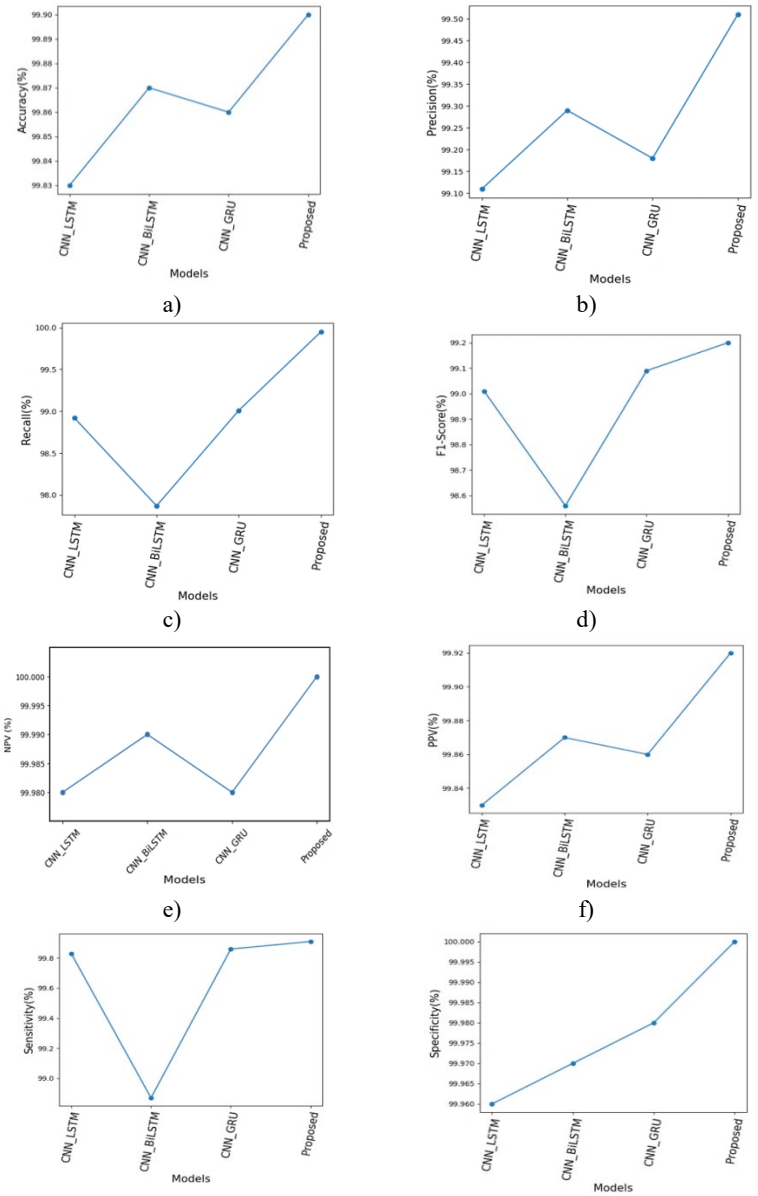


Fig. 6. Comparison Results: a) Accuracy; b) Precision; c) Recall; d) F1-score; e) NPV; f) PPV; g) Sensitivity; h) Specificity

Table 3. Performance of the system multiclass classification and error rate variations

Algorithm	Accuracy (%)	Error Rate
CNN_LSTM[39]	99.83	0.092
CNN_BiLSTM[39]	87.99	0.1
CNN_GRU[39]	86.99	0.016
Proposed	99.90	0.012

As a result, it is proposed to balance the dataset. To address this issue, the oversampling approach was utilized to balance the datasets. Synthetic samples for the minority class are generated using SMOTE-Tomek for regional expertise instead of undefined facts regarding the faction category. The model successfully reflects the dimensional and secular connection of normal spotting challenges. The proposed methodology can be used to detect and evaluate anomalies in a wide range of IoT applications and data. Thus, CNN_XGB is capable of dealing with huge amounts of data which performs superior when dealing with huge quantities of information.

5. Conclusion. Through the creation of an IDS, this research article proposed a creative approach to improve the security of IoT environments. The observation is carried out on the IoT-23 dataset, and the results show that the proposed technique achieves good performance. The findings of the proposed work show that this combined learning technique with a balanced high-performing feature selection method, SMOTE-Tomek, CNN_XGBoost, and Adam Optimizer achieved a high accuracy of 99.90%. The integration of the proposed model contributes to constructing a strong and scalable IDS that can be applied to various IoT scenarios. Thus, our research work has practical applications and paves the way for further innovation in IoT security, ultimately contributing to the growth of secure and resilient IoT ecosystems. Future work should integrate the IDS with SIEM (Security Information and Event Management) solutions to give a more comprehensive security ecosystem for IoT networks. This can enhance the system's ability to correlate events and provide a holistic view of security.

References

1. Chopra K., Gupta K., Lambora A. Future internet: The internet of things-a literature review. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, 2019. pp. 135–139.
2. Apostol I., Preda M., Nila C., Bica I. IoT botnet anomaly detection using unsupervised deep learning. Electronics. 2021. vol. 10(16). DOI: 10.3390/electronics10161876.

3. Raghuvanshi A., Singh U.K. WITHDRAWN: Internet of Things for smart cities-security issues and challenges. 2020. DOI: 10.1016/j.matpr.2020.10.849.
4. Lokhande M.P., Patil D.D., Patil L.V., Shabaz M. Machine-to-machine communication for device identification and classification in secure telerobotics surgery. Security and communication networks. 2021. no. 1. pp. 1–16. DOI: 10.1155/2021/5287514.
5. Butun I., Osterberg P., Song H. Security of the Internet of Things: Vulnerabilities, attacks, and countermeasures. IEEE Communications Surveys and Tutorials. 2019. vol. 22(1). pp. 616–644.
6. Zahra S.R., Chishti M.A. Ransomware and internet of things: A new security nightmare. In 2019 9th international conference on cloud computing, data science & engineering (confluence). IEEE, 2019. pp. 551–555.
7. Makhdoom I., Abolhasan M., Lipman J., Liu R.P., Ni W. Anatomy of threats to the internet of things. IEEE communications surveys and tutorials. 2018. vol. 21(2). pp. 1636–1675.
8. Liang L., Zheng K., Sheng Q., Huang X. A denial of service attack method for an IoT system. In 8th international conference on Information Technology in Medicine and Education (ITME). IEEE, 2016. pp. 360–364.
9. Gray C., Ayre R., Hinton K., Tucker R.S. Power consumption of IoT access network technologies. In IEEE International Conference on Communication Workshop (ICCW). IEEE, 2015. pp. 2818–2823.
10. Gormuş S., Aydın H., Ulutaş G. Security for the internet of things: a survey of existing mechanisms, protocols and open research issues. Journal of the Faculty of Engineering and Architecture of Gazi University. 2018. vol. 33(4). pp. 1247–1272.
11. Carracedo J.M., Milliken M., Chouhan P.K., Scotney B., Lin Z., Sajjad A., Shackleton M. Cryptography for security in IoT. In Fifth International Conference on Internet of Things: Systems, Management and Security. IEEE, 2018. pp. 23–30.
12. Karati A., Fan C.I., Hsu R.H. Provably secure and generalized signcryption with public verifiability for secure data transmission between resource-constrained IoT devices. IEEE Internet of Things Journal. 2019. vol. 6(6). pp. 10431–10440.
13. Fang D., Qian Y., Hu R.Q. A flexible and efficient authentication and secure data transmission scheme for IoT applications. IEEE Internet of Things Journal. 2020. vol. 7(4). pp. 3474–3484.
14. Chaabouni N., Mosbah M., Zemmari A., Sauvignac C., Faruki P. Network intrusion detection for IoT security based on learning techniques. IEEE Communications Surveys and Tutorials. 2019. vol. 21(3). pp. 2671–2701.
15. Aldweesh A., Derhab A., Emam A.Z. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems. 2020. vol. 189(5). DOI: 10.1016/j.knosys.2019.105124.
16. Albulayhi K., Sheldon F.T. An adaptive deep-ensemble anomaly-based intrusion detection system for the internet of things. In 2021 IEEE World AI IoT Congress (AIIoT). IEEE, 2021. pp. 0187–0196.
17. Alrubayyi H., Goteng G., Jaber M., Kelly J. Challenges of malware detection in the IoT and a review of artificial immune system approaches. Journal of Sensor and Actuator Networks. 2021. vol. 10(4). DOI: 10.3390/jsan10040061.
18. Al-Turaiki I., Altwaijry N. A convolutional neural network for improved anomaly-based network intrusion detection. Big Data. 2021. vol. 9(3). pp. 233–252.
19. Lam N.T. Detecting unauthorized network intrusion based on network traffic using behavior analysis techniques. International Journal of Advanced Computer Science and Applications. 2021. vol. 12(4). DOI: 10.14569/IJACSA.2021.0120407.
20. Aljumah A. IoT-based intrusion detection system using convolution neural networks. PeerJ Computer Science. 2021. vol. 7. DOI: 10.7717/peerj-cs.721.

21. Akhtar M.S., Feng T. Deep learning-based framework for the detection of cyberattack using feature engineering. Security and Communication Networks, 2021. no. 1. DOI: 10.1155/2021/6129210.
22. Liu C., Gu Z., Wang J. A hybrid intrusion detection system based on scalable K-means+ random forest and deep learning. IEEE Access. 2021. vol. 9. pp. 75729–75740.
23. Thilagam T., Aruna R. Intrusion detection for network based cloud computing by custom RC-NN and optimization. ICT Express. 2021. vol. 7(4). pp. 512–520.
24. Kanna P.R., Santhi P. Unified deep learning approach for efficient intrusion detection system using integrated spatial-temporal features. Knowledge-Based Systems. 2021. vol. 226. DOI: 10.1016/j.knosys.2021.107132.
25. Yin S.L., Zhang X.L., Liu S. Intrusion detection for capsule networks based on dual routing mechanism. Computer Networks. 2021. vol. 197. DOI: 10.1016/j.knosys.2021.107132.
26. Khan A.S., Ahmad Z., Abdullah J., Ahmad F. A spectrogram image-based network anomaly detection system using deep convolutional neural network. IEEE access. 2021. vol. 9. pp. 87079–87093.
27. Chen Y., Lin Q., Wei W., Ji J., Wong K.C., Coello C.A.C. Intrusion detection using multi-objective evolutionary convolutional neural network for Internet of Things in Fog computing. Knowledge-Based Systems. 2022. vol. 244. DOI: 10.1016/j.knosys.2022.108505.
28. Dahou A., Abd Elaziz M., Chelloug S.A., Awadallah M.A., Al-Betar M.A., Al-Qaness M.A., Forestiero A. 2022. Intrusion detection system for IoT based on deep learning and modified reptile search algorithm. Computational Intelligence and Neuroscience. 2022. no. 1. DOI: 10.1155/2022/6473507.
29. Haq M.A., Rahim Khan M.A., AL-Harbi T. Development of PCCNN-based network intrusion detection system for EDGE computing. Computers, Materials and Continua. 2022. vol. 71(1). DOI: 10.32604/cmc.2022.018708.
30. Albulayhi K., Abu Al-Haija Q., Alsuhibany S.A., Jillepalli A.A., Ashrafuzzaman M., Sheldon F.T. IoT intrusion detection using machine learning with a novel high performing feature selection method. Applied Sciences. 2022. vol. 12(10). DOI: 10.3390/app12105015.
31. Stoyanova M., Nikoloudakis Y., Panagiotakis S., Pallis E., Markakis E.K. A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues. IEEE Communications Surveys and Tutorials. 2020. vol. 22(2). pp. 1191–1221.
32. Henderi H., Wahyuningsih T., Rahwanto E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems. 2021. vol. 4(1). pp. 13–20.
33. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002. vol. 16. pp. 321–357.
34. Allam M., Nandhini M. Optimal feature selection using binary teaching learning based optimization algorithm. Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34(2). pp. 329–341.
35. Smys S., Basar A., Wang H. Hybrid intrusion detection system for internet of things (IoT). Journal of ISMAC 2020. vol. 2(04). pp. 190–199.
36. Raichura M., Chothani N., Patel D. Efficient CNN-XGBoost technique for classification of power transformer internal faults against various abnormal conditions. IET Generation, Transmission and Distribution. 2021. vol. 15(5). pp. 972–985.

37. Ullah I., Mahmoud Q.H. Design and development of RNN anomaly detection model for IoT networks. IEEE Access. 2022. vol. 10. pp. 62722–62750.
38. Susilo B., Sari R.F. Intrusion detection in IoT networks using deep learning algorithm. Information. 2020. vol. 11(5). DOI: 10.3390/info11050279.
39. Soliman S., Oudah W., Aljuhani A. Deep learning-based intrusion detection approach for securing industrial Internet of Things. Alexandria Engineering Journal. 2023. vol. 81. pp. 371–383.
40. Khelil H., Brahimi M. Toward an efficient web service composition based on an improved BTLBO algorithm. The Journal of Supercomputing. 2024. vol. 80(7). pp. 8592–8613.
41. Ullah I., Mahmoud Q.H. A framework for anomaly detection in IoT networks using conditional generative adversarial networks. IEEE Access. 2021. vol. 9. pp. 165907–165931.
42. Khuat T.T., Le M.H. Binary teaching–learning-based optimization algorithm with a new update mechanism for sample subset optimization in software defect prediction. Soft Computing. 2019. vol. 23(20). pp. 9919–9935.
43. Nazir A., He J., Zhu N., Qureshi S.S., Qureshi S.U., Ullah F., Wajahat A., Pathan M.S. (2024). A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem. Ain Shams Engineering Journal. 2024. vol. 15. no. 7. DOI: 10.1016/j.asej.2024.102777.
44. Gao X., Jamil N., Ramli M.I., Ariffin S.M.Z.S.Z. A Comparative Analysis of Combination of CNN-Based Models with Ensemble Learning on Imbalanced Data. JOIV: International Journal on Informatics Visualization. 2024. vol. 8. no. 1. pp. 456–464.
45. Zawaideh F.H., Al-Asad G., Swaneh G., Batainah S., Bakkar H. Intrusion Detection System for (IoI) Networks Using Convolutional Neural Network (CNN) and Xgboost Algorithm. Journal of Theoretical and Applied Information Technology. 2024. vol. 102(4). pp. 1750–1759.
46. Swana E.F., Doorsamy W., Bokoro P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors. 2022. vol. 22(9). DOI: 10.3390/s22093246.

Narayanarao Chokkapu — Research scholar, Department of computer science and engineering, GITAM School of Technology. Research interests: IoT security, deep learning. The number of publications — 3. nchokkap@gitam.in; GITAM Visakhapatnam Campus, Gandhi Nagar, Rushikonda, 530045, Visakhapatnam, Andhra Pradesh, India; office phone: +91(08912)790-501.

Mandapati Venkateswara Rao — Professor, Department of computer science and engineering, GITAM School of Technology. Research interests: robotics, cloud computing. The number of publications — 10. vmandapa@gitam.edu; GITAM Visakhapatnam Campus, Gandhi Nagar, Rushikonda, 530045, Visakhapatnam, Andhra Pradesh, India; office phone: +91(08912)790-501.

Boddu Bhaskara Rao — Associate professor, Department of computer science and engineering, GITAM School of Technology. Research interests: machine learning, data science, semantic web. The number of publications — 8. bboddu@gitam.edu; GITAM Visakhapatnam Campus, Gandhi Nagar, Rushikonda, 530045, Visakhapatnam, Andhra Pradesh, India; office phone: +91(08912)790-501.

УДК 004 **DOI** 10.15622/ia.23.6.10

Ч. НАРАЯНАРАО, В. МАНДАПАТИ, Б. БОДДУ

СИНЕРГЕТИЧЕСКИЕ ПОДХОДЫ К УЛУЧШЕНИЮ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ В ИНТЕРНЕТ ВЕЩЕЙ (IOT): БАЛАНСИРОВКА ХАРАКТЕРИСТИК С ПОМОЩЬЮ КОМБИНИРОВАННОГО ОБУЧЕНИЯ

Нараянарао Ч., Мандапати В., Бодду Б. **Синергетические подходы к улучшению обнаружения вторжений в Интернет вещей (IoT): балансировка характеристик с помощью комбинированного обучения.**

Аннотация. Интернет вещей (IoT) играет важную роль в обеспечении безопасности, предотвращая несанкционированный доступ, заражения вредоносным ПО и злонамеренные действия. IoT отслеживает сетевой трафик, а также поведение устройств для выявления потенциальных угроз и принятия соответствующих мер противодействия. Тем не менее, существует потребность в системе обнаружения вторжений (IDS) IoT с улучшенными возможностями обобщения, использующей глубокое обучение и передовые методы обнаружения аномалий. В этом исследовании представлен инновационный подход к IoT IDS, который сочетает в себе SMOTE-Tomek и BTLBO, CNN с XGB классификатором, который направлен на устранение дисбаланса данных, повышение производительности модели, снижение количества неправильных классификаций и улучшение общего качества набора данных. Предложенная система обнаружения вторжений IoT, используя набор данных IoT-23, достигает 99,90% точности и низкого уровня ошибок, требуя при этом существенно меньше времени выполнения. Эта работа представляет собой значительный шаг вперед в области безопасности IoT, предлагая надежное и эффективное решение IDS, адаптированное к меняющимся проблемам взаимосвязанного мира.

Ключевые слова: минимаксная нормализация, SMOTE-Tomek Link, алгоритм BTLBO, CNN с XGB, оптимизатор Adam.

Литература

1. Chopra K., Gupta K., Lambora A. Future internet: The internet of things-a literature review. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, 2019. pp. 135–139.
2. Apostol I., Preda M., Nila C., Bica I. IoT botnet anomaly detection using unsupervised deep learning. Electronics. 2021. vol. 10(16). DOI: 10.3390/electronics10161876.
3. Raghuvanshi A., Singh U.K. WITHDRAWN: Internet of Things for smart cities-security issues and challenges. 2020. DOI: 10.1016/j.matpr.2020.10.849.
4. Lokhande M.P., Patil D.D., Patil L.V., Shabaz M. Machine-to-machine communication for device identification and classification in secure telerobotics surgery. Security and communication networks. 2021. no. 1. pp. 1–16. DOI: 10.1155/2021/5287514.
5. Butun I., Osterberg P., Song H. Security of the Internet of Things: Vulnerabilities, attacks, and countermeasures. IEEE Communications Surveys and Tutorials. 2019. vol. 22(1). pp. 616–644.
6. Zahra S.R., Chishti M.A. Ransomware and internet of things: A new security nightmare. In 2019 9th international conference on cloud computing, data science & engineering (confluence). IEEE, 2019. pp. 551–555.

7. Makhdoom I., Abolhasan M., Lipman J., Liu R.P., Ni W. Anatomy of threats to the internet of things. IEEE communications surveys and tutorials. 2018. vol. 21(2). pp. 1636–1675.
8. Liang L., Zheng K., Sheng Q., Huang X. A denial of service attack method for an IoT system. In 8th international conference on Information Technology in Medicine and Education (ITME). IEEE, 2016. pp. 360–364.
9. Gray C., Ayre R., Hinton K., Tucker R.S. Power consumption of IoT access network technologies. In IEEE International Conference on Communication Workshop (ICCW). IEEE, 2015. pp. 2818–2823.
10. Gormuş S., Aydın H., Ulutaş G. Security for the internet of things: a survey of existing mechanisms, protocols and open research issues. Journal of the Faculty of Engineering and Architecture of Gazi University. 2018. vol. 33(4). pp. 1247–1272.
11. Carracedo J.M., Milliken M., Chouhan P.K., Scotney B., Lin Z., Sajjad A., Shackleton M. Cryptography for security in IoT. In Fifth International Conference on Internet of Things: Systems, Management and Security. IEEE, 2018. pp. 23–30.
12. Karati A., Fan C.I., Hsu R.H. Provably secure and generalized signcryption with public verifiability for secure data transmission between resource-constrained IoT devices. IEEE Internet of Things Journal. 2019. vol. 6(6). pp. 10431–10440.
13. Fang D., Qian Y., Hu R.Q. A flexible and efficient authentication and secure data transmission scheme for IoT applications. IEEE Internet of Things Journal. 2020. vol. 7(4). pp. 3474–3484.
14. Chaabouni N., Mosbah M., Zemmari A., Sauvignac C., Faruki P. Network intrusion detection for IoT security based on learning techniques. IEEE Communications Surveys and Tutorials. 2019. vol. 21(3). pp. 2671–2701.
15. Aldweesh A., Derhab A., Emam A.Z. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems. 2020. vol. 189(5). DOI: 10.1016/j.knosys.2019.105124.
16. Albulayhi K., Sheldon F.T. An adaptive deep-ensemble anomaly-based intrusion detection system for the internet of things. In 2021 IEEE World AI IoT Congress (AIIoT). IEEE, 2021. pp. 0187–0196.
17. Alrubayyi H., Goteng G., Jaber M., Kelly J. Challenges of malware detection in the IoT and a review of artificial immune system approaches. Journal of Sensor and Actuator Networks. 2021. vol. 10(4). DOI: 10.3390/jsan10040061.
18. Al-Turaiki I., Altwaijry N. A convolutional neural network for improved anomaly-based network intrusion detection. Big Data. 2021. vol. 9(3). pp. 233–252.
19. Lam N.T. Detecting unauthorized network intrusion based on network traffic using behavior analysis techniques. International Journal of Advanced Computer Science and Applications. 2021. vol. 12(4). DOI: 10.14569/IJACSA.2021.0120407.
20. Aljumah A. IoT-based intrusion detection system using convolution neural networks. PeerJ Computer Science. 2021. vol. 7. DOI: 10.7717/peerj-cs.721.
21. Akhtar M.S., Feng T. Deep learning-based framework for the detection of cyberattack using feature engineering. Security and Communication Networks, 2021. no. 1. DOI: 10.1155/2021/6129210.
22. Liu C., Gu Z., Wang J. A hybrid intrusion detection system based on scalable K-means+ random forest and deep learning. IEEE Access. 2021. vol. 9. pp. 75729–75740.
23. Thilagam T., Aruna R. Intrusion detection for network based cloud computing by custom RC-NN and optimization. ICT Express. 2021. vol. 7(4). pp. 512–520.
24. Kanna P.R., Santhi P. Unified deep learning approach for efficient intrusion detection system using integrated spatial-temporal features. Knowledge-Based Systems. 2021. vol. 226. DOI: 10.1016/j.knosys.2021.107132.

25. Yin S.L., Zhang X.L., Liu S. Intrusion detection for capsule networks based on dual routing mechanism. Computer Networks. 2021. vol. 197. DOI: 10.1016/j.knosys.2021.107132.
26. Khan A.S., Ahmad Z., Abdullah J., Ahmad F. A spectrogram image-based network anomaly detection system using deep convolutional neural network. IEEE access. 2021. vol. 9. pp. 87079–87093.
27. Chen Y., Lin Q., Wei W., Ji J., Wong K.C., Coello C.A.C. Intrusion detection using multi-objective evolutionary convolutional neural network for Internet of Things in Fog computing. Knowledge-Based Systems. 2022. vol. 244. DOI: 10.1016/j.knosys.2022.108505.
28. Dahou A., Abd Elaziz M., Chelloug S.A., Awadallah M.A., Al-Betar M.A., Al-Qaness M.A., Forestiero A. 2022. Intrusion detection system for IoT based on deep learning and modified reptile search algorithm. Computational Intelligence and Neuroscience. 2022. no. 1. DOI: 10.1155/2022/6473507.
29. Haq M.A., Rahim Khan M.A., AL-Harbi T. Development of PCCNN-based network intrusion detection system for EDGE computing. Computers, Materials and Continua. 2022. vol. 71(1). DOI: 10.32604/cmc.2022.018708.
30. Albulayhi K., Abu Al-Haija Q., Alsuhibany S.A., Jillepalli A.A., Ashrafuzzaman M., Sheldon F.T. IoT intrusion detection using machine learning with a novel high performing feature selection method. Applied Sciences. 2022. vol. 12(10). DOI: 10.3390/app12105015.
31. Stoyanova M., Nikoloudakis Y., Panagiotakis S., Pallis E., Markakis E.K. A survey on the internet of things (IoT) forensics: challenges, approaches, and open issues. IEEE Communications Surveys and Tutorials. 2020. vol. 22(2). pp. 1191–1221.
32. Henderi H., Wahyuningsih T., Rahwanto E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems. 2021. vol. 4(1). pp. 13–20.
33. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002. vol. 16. pp. 321–357.
34. Allam M., Nandhini M. Optimal feature selection using binary teaching learning based optimization algorithm. Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34(2). pp. 329–341.
35. Smys S., Basar A., Wang H. Hybrid intrusion detection system for internet of things (IoT). Journal of ISMAC 2020. vol. 2(04). pp. 190–199.
36. Raichura M., Chothani N., Patel D. Efficient CNN-XGBoost technique for classification of power transformer internal faults against various abnormal conditions. IET Generation, Transmission and Distribution. 2021. vol. 15(5). pp. 972–985.
37. Ullah I., Mahmoud Q.H. Design and development of RNN anomaly detection model for IoT networks. IEEE Access. 2022. vol. 10. pp. 62722–62750.
38. Susilo B., Sari R.F. Intrusion detection in IoT networks using deep learning algorithm. Information. 2020. vol. 11(5). DOI: 10.3390/info11050279.
39. Soliman S., Oudah W., Aljuhani A. Deep learning-based intrusion detection approach for securing industrial Internet of Things. Alexandria Engineering Journal. 2023. vol. 81. pp. 371–383.
40. Khelil H., Brahimi M. Toward an efficient web service composition based on an improved BTLBO algorithm. The Journal of Supercomputing. 2024. vol. 80(7). pp. 8592–8613.

41. Ullah I., Mahmoud Q.H. A framework for anomaly detection in IoT networks using conditional generative adversarial networks. IEEE Access. 2021. vol. 9. pp. 165907–165931.
42. Khuat T.T., Le M.H. Binary teaching–learning-based optimization algorithm with a new update mechanism for sample subset optimization in software defect prediction. Soft Computing. 2019. vol. 23(20). pp. 9919–9935.
43. Nazir A., He J., Zhu N., Qureshi S.S., Qureshi S.U., Ullah F., Wajahat A., Pathan M.S. (2024). A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem. Ain Shams Engineering Journal. 2024. vol. 15. no. 7. DOI: 10.1016/j.asej.2024.102777.
44. Gao X., Jamil N., Ramli M.I., Ariffin S.M.Z.S.Z. A Comparative Analysis of Combination of CNN-Based Models with Ensemble Learning on Imbalanced Data. JOIV: International Journal on Informatics Visualization. 2024. vol. 8. no. 1. pp. 456–464.
45. Zawaideh F.H., Al-Asad G., Swaneh G., Batainah S., Bakkar H. Intrusion Detection System for (IoI) Networks Using Convolutional Neural Network (CNN) and Xgboost Algorithm. Journal of Theoretical and Applied Information Technology. 2024. vol. 102(4). pp. 1750–1759.
46. Swana E.F., Doorsamy W., Bokoro P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors. 2022. vol. 22(9). DOI: 10.3390/s22093246.

Нараянарао Чоккапу — научный сотрудник, кафедра компьютерных наук и инженерии, Технологическая школа GITAM. Область научных интересов: безопасность интернета вещей, глубокое обучение. Число научных публикаций — 3. nchokkap@gitam.in; кампус GITAM Вишакхапатнам, Ганди Нагар, Рушиконда, 530045, Вишакхапатнам, Андхра-Прадеш, Индия; р.т.: +91(08912)790-501.

Мандапати Венкатесвара Рао — профессор, кафедра компьютерных наук и инженерии, Технологическая школа GITAM. Область научных интересов: робототехника, облачные вычисления. Число научных публикаций — 10. vmandapa@gitam.edu; кампус GITAM Вишакхапатнам, Ганди Нагар, Рушиконда, 530045, Вишакхапатнам, Андхра-Прадеш, Индия; р.т.: +91(08912)790-501.

Бодду Бхаскара Рао — доцент, кафедра компьютерных наук и инженерии, Технологическая школа GITAM. Область научных интересов: машинное обучение, наука о данных, семантическая сеть. Число научных публикаций — 8. bboddu@gitam.edu; кампус GITAM Вишакхапатнам, Ганди Нагар, Рушиконда, 530045, Вишакхапатнам, Андхра-Прадеш, Индия; р.т.: +91(08912)790-501.

DOI 10.15622/ia.23.6.11

H. DONG

CONVOLUTIONAL-FREE MALWARE IMAGE CLASSIFICATION USING SELF-ATTENTION MECHANISMS

Dong H. **Convolutional-free Malware Image Classification using Self-attention Mechanisms.**

Abstract. Malware analysis is a critical aspect of cybersecurity, aiming to identify and differentiate malicious software from benign programmes to protect computer systems from security threats. Despite advancements in cybersecurity measures, malware continues to pose significant risks in cyberspace, necessitating accurate and rapid analysis methods. This paper introduces an innovative approach to malware classification using image analysis, involving three key phases: converting operation codes into RGB image data, employing a Generative Adversarial Network (GAN) for synthetic oversampling, and utilising a simplified Vision Transformer (ViT)-based classifier for image analysis. The method enhances feature richness and explainability through visual imagery data and addresses imbalanced classification using GAN-based oversampling techniques. The proposed framework combines the strengths of convolutional autoencoders, hybrid classifiers, and adapted ViT models to achieve a balance between accuracy and computational efficiency. As shown in the experiments, our convolutional-free approach possesses excellent accuracy and precision compared with convolutional models and outperforms CNN models on two datasets, thanks to the multi-head attention mechanism. On the Big2015 dataset, our model outperforms other CNN models with an accuracy of 0.8369 and an AUC of 0.9791. Specifically, our model reaches an accuracy of 0.9697 and an F1 score of 0.9702 on MALIMG, which is extraordinary.

Keywords: malware classification, cybersecurity, deep learning, transformer.

1. Introduction. Malware analysis is essential for identifying malicious software on a host system and distinguishing it from benign programs. Despite significant advancements in cybersecurity measures, malware remains a potent hazard in cyberspace. Given the challenges posed by malware, accurate classification techniques with efficient computation are vital for safeguarding computers against infection or effectively removing malware. The field of malware analysis faces two primary challenges: the swift development and distribution of malware and the use of sophisticated evasion techniques. The internet has accelerated the development and dissemination of new malware, creating a vulnerability that can infect numerous systems despite antivirus defences [1]. To meet the ever-evolving threats of cyberattacks, recent advancements in research focus on real-time ransomware detection and combining machine learning techniques in autonomous protection to combat malware development and distribution and counter sophisticated evasion techniques.

Understanding malware attacks and their methodologies is crucial for developing robust defences against such threats, and this logic is utilised in much of the existing research. Malware files can be analysed using static and

dynamic approaches. Static analysis involves examining the code without executing the file, enabling quick diagnosis [2, 3], while dynamic analysis scrutinises the system and network behaviour of malicious files in a controlled environment for a more thorough analysis [2, 4]. A newer method, in-memory analysis, involves analysing memory snapshots of active processes to detect unusual process activation consistently [4, 5]. In the domain of malware analysis, numerous techniques have been deployed; traditional methods like signature-based detection, string analysis, and malware binaries hashing still retain a certain level of effectiveness. However, the sophisticated nature of malware necessitates advanced detection strategies, as traditional ML methods often struggle with high-dimensional data and extracting intricate patterns. A recent survey [6] highlights that DL often outperforms traditional ML techniques in malware detection and analysis tasks, attributed to its automatic feature extraction, capacity to learn non-linear relationships, and prowess in image analysis. DL has long been employed in network intrusion detection [7, 8] and emerged as a viable option for extracting insights in malware analysis, with convolutional neural networks (CNN) gaining popularity due to their strong capability in feature representation learning and working with imagery inputs [2, 4, 9]. The progress in contemporary malware analysis research is indeed valuable; however, the development of malware classification techniques with better accuracy, fast computation, and generalisability is worth researching. Overall, effective malware classification should provide highly accurate performance and scalability to protect against network-wide attacks in various system environments, as well as demonstrate adaptability to evolving threat landscapes.

This paper introduces an innovative malware classification method through image analysis, comprising three phases: translating operation codes into RGB image data, deploying a Generative Adversarial Network (GAN) for adaptive synthetic oversampling, and implementing a self-attention block-based classifier for image analysis. Initially, the binary files of analysed programmes are transformed into RGB image data for imagery analysis. The adoption of the 3-channel RGB format is justified due to its widespread prevalence as the standard colour model in DL frameworks. Pre-trained CNNs are typically trained on large-scale datasets like ImageNet, and this standard enables seamless integration of pre-existing models without extensive modifications, as demonstrated by the practicality of the transfer learning approach in this context, which simplifies implementation and fine-tuning processes. To tackle the challenge of imbalanced software sample classification, we utilise GAN-based oversampling techniques to create synthetic samples for minority classes. Within this framework, the generator is designed as a

convolutional autoencoder (AE) with denoising capabilities, pre-trained on the comprehensive dataset; conversely, the discriminator functions as a hybrid classifier to discern the "fake" samples produced by the generator; this strategy mitigates the effects of imbalanced data on model performance. Finally, we adapt the self-attention blocks for image classification. Instead of employing the original ViT model with 16 transformer blocks, we adjust its structure and depth to achieve an equilibrium between accuracy and computational efficiency. The experiments demonstrate that the light transformer model outperforms all CNN models in malware classification challenges, with an accuracy of 0.80 on Big2015 and 0.966 on MALIMG, respectively.

The novelty of the proposed methodology lies in its multifaceted approach to enhancing malware classification via advanced image analysis techniques. First, the integration of GAN for adaptive synthetic oversampling is particularly noteworthy; it not only alleviates the prevalent issue of class imbalance in malware training datasets but does so by generating high-quality, diverse synthetic samples that improve the robustness of the classifier. The use of AE as the generator with denoising capabilities refines this process by ensuring that the synthetic malware images contribute positively to model training without introducing noise. Secondly, the strategic modification of the self-attention block-based classifier by optimising the ViT model structure showcases an advanced approach with both accuracy and efficiency. This tailored adaptation offers a scalable, efficient solution to contemporary cybersecurity challenges.

The paper is structured as follows: Section 2 provides a review of the existing literature on malware analysis, mainly image-based techniques. Section 3 outlines the proposed workflow and techniques. Section 4 details the experimental setup, results, and discussions. Finally, Section 5 offers conclusions and discusses potential future research directions.

2. Related Work. Recent advancements in data mining, machine learning, and deep learning algorithms have significantly enhanced the effectiveness of malware analysis. While various deep learning (DL) models can be utilised for security domains, there is a growing trend towards the application of AE. In study [10] the authors employed an AE model for network-based anomaly intrusion detection and malware classification, aiming to improve performance across different evaluation metrics. Paper [11] conducted an analysis of Android malware using image classification, employing AE with three distinct structures: a feed forward network, CNN, and VGG19, for representation learning. The experimental results underscore the exceptional representation extraction capabilities of CNN-based models. In paper [12] the authors introduced a hybrid DL approach that combines

CNN with bidirectional LSTM for malware detection. Study [13] integrated attention mechanisms with CNN to direct the model's focus towards regions of higher importance for the classification task during the learning process. Paper [9] developed a malware detection framework using Depthwise Efficient Attention Module and DenseNet, using spatial pyramid pooling to improve detection performance and overcome obfuscation sensitivity and computational overhead. The attention-based approach in these models is built upon convolutional layers. Despite the promising results in the referenced paper, there are limitations associated with using convolutional-based attention mechanisms. CNN may struggle to learn representations from long-range dependencies due to the constrained receptive field of convolutional layers, potentially hindering the model's ability to grasp global context and relationships among distant elements in the input sequence.

Recently, transfer learning (TL) has emerged as a prevalent methodology within security research. This technique, by applying knowledge gained from addressing one issue to a similar and related one, can mitigate the issue of inadequate training data and enhance the capability to detect malware or network attacks with great robustness. A notable application of TL involves an automated vulnerability detection method that converts source code into a minimal intermediate representation, employing pre-trained convolutional classifiers for analysis, demonstrating high granularity [14]. Research employing TL frequently favours CNN-based models, not only for malware but also for conducting network traffic analysis. Study [15] introduced a ConvNet model that employs transfer learning for network intrusion detection, markedly improving the detection accuracy for both known and novel attacks, as verified through experiments on the NSL-KDD dataset. Additionally, a ConvNet-based malware detection model, through a novel framework that incorporates a deep unsupervised pre-training clustering technique, surpassed the performance of ConvNets with a shallower structure [16]. TL-based approaches often necessitate deep models, especially for malware imagery analysis. While a shallow CNN model with merely three convolutional layers and one feed forward layer may suffice for network intrusion detection due to fewer features [17], opcode-based malware images typically demand more layers for effective representation learning. In the Malbert framework [18], a deep model comprising twelve encoder blocks for representation learning, a pre-classifier layer for anomaly detection, and a malware classification layer were used. Despite its complex structure, Malbert surpassed other deep learning models, such as LSTM, and ensemble learning models, such as Random Forests. A study comparing pre-trained CNN models for malware classification underscored the efficacy of TL and examined prevalent challenges like

over-fitting and high resource consumption, utilising simpler CNN models like MobileNet and ResNet50 [19]. Ultimately, the study revealed the importance of employing deep and complex models, where MobileNet showed more stable outcomes than ResNet50, yet with fewer parameters, indicating a potential equilibrium between classification efficacy and resource efficiency.

The existing literature underscores the effectiveness of CNNs and AEs in security domains, particularly malware analysis capabilities. Furthermore, the adoption of TL strategies, leveraging data from related fields, has been a cornerstone in advancing security solutions. Building upon the remarkable progress spotlighted in recent studies on malware detection through DL techniques, this paper proposes a novel malware analysis framework integrating self-attention mechanisms and GAN oversamplers with convolutional AE as generators. This research aims to utilise the inherent strengths of ViT self-attention frameworks while implementing a relatively lightweight model structure. Moreover, by employing convolutional AEs within GANs to generate oversampled data, this study seeks to address the challenges of imbalanced datasets, a recurrent issue in malware classification tasks.

3. Methodology. Figure 1 illustrates the overall workflow of the proposed method, which comprises three major steps: malware conversion, GAN-based oversampling, and image classification. Transforming malware binaries into images is a common initial step in imagery-based malware analysis approaches, aimed at visually representing binary data for pattern detection using image analysis techniques. Malware files in any binary PE format will have each byte read as an 8-bit unsigned integer before being organised into an imagery array for further processing. Starting with the original malware samples, the binary files are opened in binary reading mode. Each byte of the binary data is then converted into its hexadecimal representation. Subsequently, the hexadecimal values are used to create images: in greyscale images, each hexadecimal value corresponds directly to a pixel value, where 0x00 represents black and 0xFF represents white, with intermediate values translating to shades of grey; in RGB images, the hexadecimal data is distributed across the three-colour channels (Red, Green, Blue). Finally, the numeric array dimensions are reshaped to the desired image array size, typically from 128*128 to 256*256 pixels per channel.

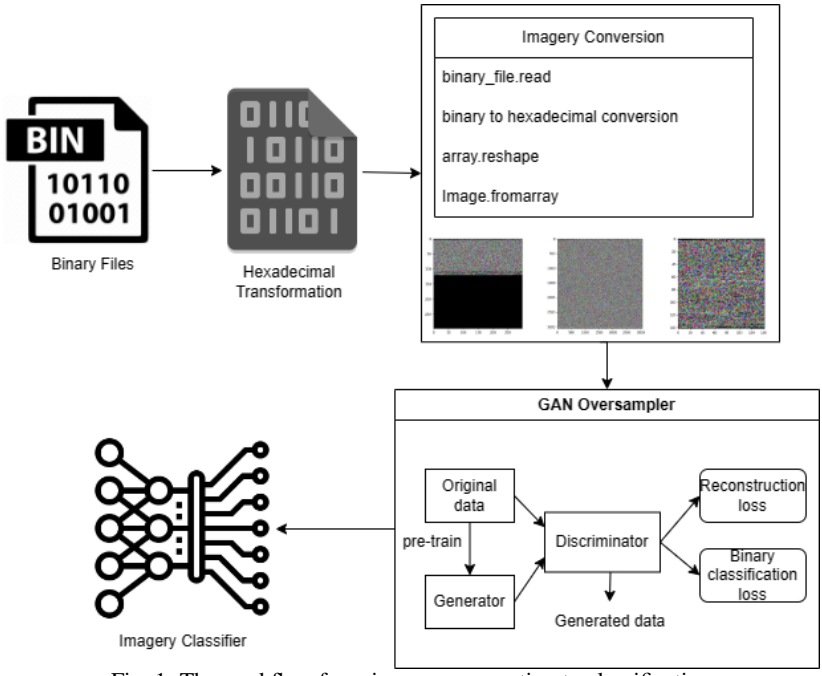


Fig. 1. The workflow from imagery generation to classification

Following data preparation, the initial step involves analysing the distribution of different classes. Since certain types of malware are relatively scarce, data imbalances emerge. To address this issue, a GAN-based oversampling technique is employed to rectify the skewed class distribution by generating synthetic samples. Lastly, a self-attention-based convolution-free image classifier is devised for the classification task. A comprehensive elaboration of these pivotal stages is provided in the following sub-sections.

3.1. GAN-based Oversampling. GANs are a class of artificial intelligence algorithms proposed by the authors in [20]. GANs consist of two neural networks, the generator and the discriminator, which are concurrently trained in a competitive manner. The generator's objective is to create data that closely resembles the original input, while the discriminator's role is to distinguish between generated and authentic data. This adversarial framework compels both networks to enhance their performance, culminating in the generator producing remarkably realistic results. GANs can acquire deep representations by propagating back-propagation signals through the competitive training of the generator (G) and discriminator (D). Through

pre-training, G learns from original input X and endeavours to generate synthetic samples X' that closely match X, hence aiming to minimise the reconstruction loss. On the other hand, D receives both generated data and original input and aims to identify fake data from the real ones, leading to the objective of minimising the loss for the binary classification task. The objective of the loss function of the GAN framework can be represented as follows:

$$\min_G \max_D \mathscr{L}(D,G) = \mathbb{E}_{x\sim P(x)}[\log D(x)] + \mathbb{E}_{z\sim Pz(z)}[\log(1-D(G(z)))], \quad (1)$$

where $P(X)$ represents the distribution of the original data and $P(z)$ is the distribution of the generator's noise input z. Function $G(z)$ maps random noise to generator models with weights learned from original data, namely learns from X and generates X'; meanwhile, function $D(x)$ represents the probability of identifying that x is real data rather than generated. During the optimisation process, the former should be minimised, and the latter should be maximised. By playing this min-max game, the generator G is forced to produce more realistic samples matching the training data distribution to fool the discriminator, allowing GAN to generate new synthetic samples for minority class enhancement.

A GAN-based oversampling technique could work well with sequential network traffic data, in which a simple model structure with only one-dimensional layers is utilised [21]. For malware imagery analysis, a more complicated model structure is essential. Figure 2 illustrates the GAN oversampler's model structures. The generator is designed as a deep convolutional AE, where both the encoder and decoder comprise two convolutional blocks. Each encoding block consists of three Conv2D layers and one Pooling2D layer. In the latent space, an additional convolutional layer is added. Subsequently, the decoder commences the reconstruction process in two blocks, each featuring three Conv2D layers and one UpSampling2D layer. The convolutional layer units are consistent within each block, with the units across the four blocks specified as "64, 128, 256 for the latent space, 128, 64," respectively. Finally, the reconstruction phase yields the output using a Conv2D layer, producing a 3-channel output representing the reconstructed image. The discriminator initiates with 128-unit initial convolutional layers, followed by batch normalisation and global average pooling. Batch normalisation aids in stabilising the model training and accelerating convergence. Global Average Pooling, instead of flattening

the multi-dimensional layers directly, reduces the total parameters and enhances computational efficiency by leveraging a globally learned parameter set. Moreover, it directs the network to prioritise crucial features, thereby improving interpretability and generalisation. Subsequently, a dropout layer is incorporated to prevent over-fitting, followed by two feed forward layers, among which the last layer is for binary classification output.

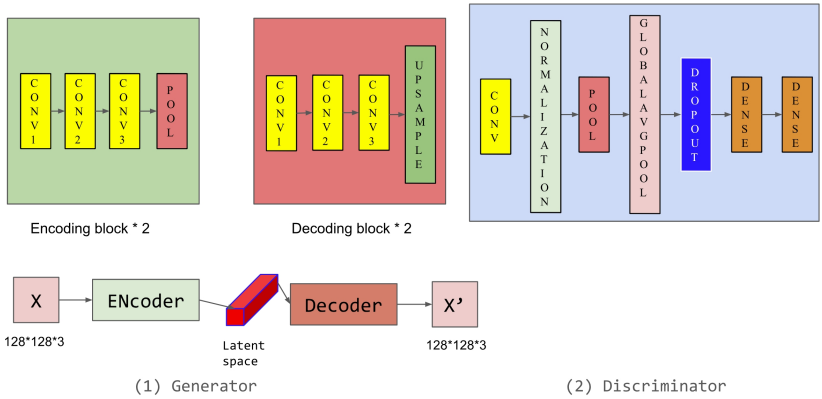


Fig. 2. The workflow from imagery generation to classification in GAN Oversampler

Algorithm 1 presents the training and oversampling process of GANOversampler, primarily focusing on enhancing the representation of minority classes in the dataset. This process generally includes GAN initialisation, a pre-training generator, and oversampling by training and applying the discriminator. The input of the oversampler takes feature sets *X_train* and label sets *y_train* as input. During initialisation, the default minority class threshold is set at 1000, which is adjustable. This quantification is selected based on the observations across selected experimental datasets, in which a threshold of 1000 sufficiently equalises the representation of minority samples. In the oversampling phase, the discriminator differentiates samples by assessing their predicted probabilities. A threshold of 0.5 is employed as a critical distinction, where samples above this threshold are deemed indistinguishable from genuine data by the discriminator. These selected samples are then used in the oversampling process to enhance minority class representation. The final outputs of the algorithm are an oversampled feature set and an oversampled label set.

Algorithm 1. GANOversampler

G = $generator()$
D = $discriminator()$
define $GAN(G,D)$
decide an array of the class id of minorities $minority\_class\_id$
decide the threshold of minorities $threshold = 1000$
Prepare the input data, feature set X_train and label set y_train
Train the discriminator on the original data $G.fit(X\_train)$
for $i \in minority\_class\_id$ **do**
    Create subset filtering by minority id $X(i)$, $y(i)$
    Count the current volume of i class $cnt = len(y(i))$
    while $cnt < threshold$ **do**
        Generate synthetic samples $sample\_raw = G.predict(noise)$
        Keep only those samples predicted to be True $new\_samples = sample\_raw[D.predict(sample\_raw) > 0.5]$
        Add $new\_samples$ into X_train and y_train and update cnt
    end while
end for
Shuffle the updated X_train and y_train

The generator learns the complex distributions and generates new data instances, while the discriminator evaluates their similarity to real data. This adversarial process continuously improves the quality of synthetic samples to closely mimic real data characteristics, demonstrating the generator's ability to learn and replicate the target distribution. With a customised threshold indicating the preferred volume of minorities, the oversampler can generate synthetic samples of under-represented classes from original volumes to the desired one, thereby balancing the dataset and providing equal training opportunities for all classes. Among traditional oversampling methods, simple random oversampling only copies the original samples. K-nearest-neighbour (KNN)-based methods like the Synthetic Minority Oversampling Technique (SMOTE) and the Adaptive Synthetic (ADASYN) fall short when working with high-dimensional data and can be time-consuming due to the computation of KNN. GANs, on the other hand, can navigate these high-dimensional spaces to produce more accurate and feasible synthetic data. The model is trained extensively before performing oversampling; hence, it will not be as time-consuming as KNN-based oversampling methods. Overall, the introduced oversampling technique excels not only in enhancing the sample size but also in preserving the quality and diversity of synthetic data, thereby ensuring that the augmented dataset supports effective model training.

3.2. Self-attention-based Classifier. As mentioned in the previous session, while convolutional-based attention mechanisms are effective for modelling local dependencies, they may struggle to effectively capture and utilise long-range relationships across the input data. On the other hand, self-attention mechanisms, as employed in ViT, allow for the modelling of interactions between all input elements simultaneously without being constrained by fixed receptive fields. This enables ViT to capture long-range dependencies more effectively and facilitates better integration of global context into the model's representations.

Our study suggests utilising the self-attention mechanism for the classification of malware imagery. This mechanism was initially introduced in the transformer model for machine translation [22], comprising an encoder and a decoder. The encoder consists of stacked self-attention and fully connected layers, while the decoder integrates multi-head attention over the encoder output. The attention mechanism employed is known as Scaled Dot-Product Attention, which calculates dot products of queries and keys, scales them, and applies a SoftMax function. With several input sets, queries, and keys with dimensions of d_k and values with a dimension of d_v, a set of queries can be packed into Q, K, and V respectively. Afterwards, the attention matrix can be calculated as follows, in which the scaling factor $1/\sqrt{d_k}$ is to counteract the potential issue of large value dot products in case of large values of d_k.

$$Attention(Q,K,V) = softmax(QK^T/\sqrt{d_k})V. \quad (2)$$

Within the classifier, the initial input image undergoes patch tokenization to divide it into multiple patches. These patches, referred to as tokens, are subsequently embedded to derive the value vectors (V). Following this, distinct linear projections are utilised on the value vectors to create query (Q) vectors and key (K) vectors, which are essential for subsequent self-attention operations. This query-key-value mechanism enables ViT models to concentrate attention on the most pertinent areas of the input by assessing the similarity between queries and keys [22]. In ViT, images are split into small patches and treated as tokens [23], which are embedded and fed to the transformer architecture. The position embedding, which is essential to retaining the imagery patches' positional information, is combined with patch embedding as input too.

Figure 3 illustrates the model architecture of the proposed malware classifier.

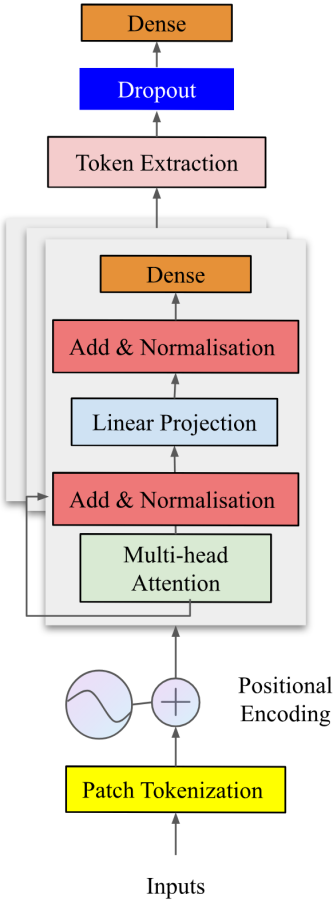


Fig. 3. The model structure of malware classifier with multiple self-attention-based blocks

Following image patching and embedding, the model incorporates multiple self-attention blocks. Within each block, multi-head attention applies separate linear projections to queries, keys, and values, enabling parallel attention computations. Each multi-head attention module contains eight attention layers in parallel, with reduced dimensions to ensure computational efficiency. Layer normalisation is applied after each block to stabilise training and mitigate issues such as vanishing gradients. Two feed forward

layers utilising GELU activation top each transformer block for non-linear transformation, and the block outputs are concatenated to represent the input image. Residual connections link stacked transformer blocks. Subsequently, a token extraction layer flattens the outputs, followed by dropout to prevent overfitting. Finally, an additional feed forward layer extracts features while the last layer classifies inputs. We optimise the block numbers based on performance, aiming to find a balance between performance and relatively low model complexity.

3.3. The Optimised Training Process. We employ several techniques to enhance stable and efficient training, including the following: TL with pre-trained weight parameters to leverage prior knowledge and avoid lengthy training from scratch; implementing class weight initialisation based on the proportion of classes in the dataset to address imbalance classification; utilising adaptive learning rate during training.

3.3.1. Transfer Learning. It is common for software to contain diverse types of malware, necessitating the classification of different malware labels. Therefore, malware image classification inherently involves numerous, imbalanced labels. Our approach pre-trains the classifier on the ImageNet dataset, which is proven beneficial for learning generalised features [24]. Initially, the base model is trained on ImageNet with a 1000-class output setting to learn sufficient parameters for complex classification tasks. Subsequently, the output layer is removed, a dropout layer is added, and feed forward layers with corresponding output units specified in the malware imagery set are incorporated. This leverages the pre-trained model for feature extraction from malware images, utilising the learned feature representations from ImageNet. However, the final layers of the pre-trained model require retraining for the new task to acquire task-specific weights. The base model layers are initially frozen by setting their trainable parameter to "false," transforming the pre-trained model into a fixed feature extractor for the new dataset. During this phase, only the classification head layers, initialised randomly, are trained to discern patterns from the extracted features, as outlined in Table 1. The model undergoes training in the frozen setting for 20 epochs. Following the training of the classification head, the base model layers become trainable by setting their trainable parameter to True. Subsequently, the complete model, encompassing both pre-trained and randomly initialised layers, undergoes end-to-end fine-tuning on the target data for 20 epochs with a reduced learning rate, as detailed in Table 1. This process facilitates further optimisation of feature representations to achieve tailored adaptation to new tasks.

Table 1. Classification Model Settings

Model Structure
InputLayer Transformer_block (stacked) LayerNormalization ExtractToken – output size(768) Dropout – rate(0.2) (Dense(256,activation="gelu", kernel_regularizer=regularizers.l2(0.01)) Dense(num_classes, activation='softmax', kernel_regularizer=regularizers.l1(0.01))
Optimistion
Stage 1: optimizers.AdamW(weight_decay=0.01, learning_rate=0.005) Stage 2: optimizers.AdamW(weight_decay=0.05, learning_rate=0.0001)

3.3.2. Weight Initialisation. To mitigate the impact of imbalanced data, in addition to employing oversampling techniques to equalise the distribution of different classes, we also incorporated class weights to prioritise the minority classes during training. The initialisation of class weights entails computing weights for each class according to their distribution in the training dataset, aiming to tackle class imbalances. Algorithm 2 outlines the key steps. The initialised class weights dictionary is then passed to the model during compilation to account for class imbalance during training.

Algorithm 2. WeightInit

Prepare input: Original label set y_train
Extract the unique class labels $unique\_classes$
Calculate the frequency of unique class labels $class\_counts$
Calculate the total number of samples $total = len(y\_train)$
for $class\_id \in unique\_classes$ **do**
    $weight(class\_id) = total/(num\_classes * count)$
end for
Output: Standardised and return the class weight array W

3.3.3. Optimisation with Adaptive Learning Rate. Adam [25], an adaptive gradient algorithm, has been widely favoured for compiling DL models. However, there are variations that can enhance its performance. According to an optimisation-focused study [26], L2 regularisation in Adam is ineffective, and weight decay is only applied after parameter updates, even

though weight decay is crucial for preventing over-fitting. This research explores both L2 regularisation and weight decay regularisation, demonstrating that the proposed Decoupled Weight Decay Regularisation (DWDR) is more effective. The main concept behind DWDR involves adding an extra term for weight decay during the parameter update step. With true label y and predicted label y', the original binary cross-entropy function can be presented as (3), while the regularisation addition to the loss function can be represented as (4):

$$BCE\_loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y^i \log y'^i + (1-y^i)\log(1-y'^i)\right), \tag{3}$$

$$loss = \text{BCE\_loss} + \frac{\lambda}{2m}\sum_{w} w^2. \tag{4}$$

In which w represents the weight learned and λ is a hyper-parameter that need to be initialised manually. When using DWDR, the weight update process from w_t to w_{t+1} is updated with a subtraction from the weights as (5), in which lr represents the learning rate and β is another hyper-parameter for weight decay. The term in the square bracket is the original updating step. In this way, we reduce the updated weight by a small portion at each step.

$$w_{t+1} = [w_t - lr \times gradient] - lr \times \beta \times w_t. \tag{5}$$

As illustrated in Table 1, we initially freeze the base model and train it with a larger learning rate and a smaller decay hyper-parameter, followed by fine-tuning it with a smaller learning rate and a larger decay hyper-parameter. In the first stage of training, this configuration accelerates the training process, enabling the model to rapidly learn general patterns and features from new data. It also enhances generalisation by allowing the model to adapt more effectively to the new dataset. During the fine-tuning stage, the smaller learning rate and larger decay contribute to stabilising the training process and preventing over-fitting. We employ this approach to maximise the benefits of adopting transfer learning.

4. Experiments

4.1. Dataset Description. To thoroughly validate the proposed methods, we conduct experiments on two malware imagery datasets within a multi-classification setting. Both datasets are divided into training, validation, and testing subsets as pre-defined by the original authors. The first dataset, Big2015 [27], was proposed by Microsoft and comprises 10,470 malware

files from nine distinct families. Each file is characterised by an identifier, a 20-character hash value, and a class label. The class labels, malware types, and their original proportions in the training set, as well as after the application of oversampling, are presented in Table 2. A backdoor is a type of malware that permits unauthorised access to a computer system, typically bypassing standard authentication mechanisms to facilitate remote control or data theft. Adware constitutes unwanted software that displays advertisements on a user's device. Obfuscated malware is malicious software with code deliberately obfuscated to evade detection by security software. A worm is a self-replicating type of malware that exploits security vulnerabilities and spreads across computer networks. A Trojan is a type of malware commonly disguised as legitimate software. Similarly, a Trojan downloader is a type of Trojan horse designed to download and install additional malware onto the infected system.

Table 2. Data Summary for Big2015

Labels	Malware Types	Initial Weight	Weight after oversampling
Gatak	Backdoor	9.3%	12.3%
Kelihos_ver1	Backdoor	3.7%	9.6%
Kelihos_ver3	Backdoor	27.1%	17.8%
Lollipop	Adware	22.8%	15.0%
Obfuscator.ACY	Any obfuscated malware	11.3%	7.4%
Ramnit	Worm	14.1%	9.3%
Simda	Backdoor	0.4%	8.0%
Tracur	TojanDownloader	6.9%	9.1%
Vundo	Tojan	4.4%	11.5%

The second dataset, MALIMG [28], was proposed in the paper for malware imagery visualisation research. It consists of 9,458 samples from 25 different malware families. The class labels and their original proportions in the training set, as well as after the application of oversampling, are presented in Table 3. Apart from the malware types covered in Big2015, MALIMG contains several new types: Dialer malware is malicious software designed to connect a system to a network or phone number for a fraudulent purpose. Rogue malware is generally characterised by deceptive behaviour; it pretends to be legitimate but can cause significant harm once installed. PWS, short for Password Stealing Ware, aims at stealing sensitive information such as login credentials, passwords, and other personal data.

The malware images contain transformed binary data, with sections such as .text holding the executable code, .rdata containing read-only data like constant values and strings, .data storing initialised data, and .rsrc housing

specific resources used by the executable and version information. Various code sections will be mapped to distinct textures within the images, as illustrated in Figure 1. Our approach relies on learning these anomalous patterns to inform the classification process by identifying similarities among the patterns.

Table 3. Data Summary for MALIMG

Labels	Malware Types	Initial Weight	Weight after oversampling
Allaple.A	Worm	31.6%	13.5%
Allaple.L	Worm	17.1%	7.3%
Yuner.A	Worm	8.6%	3.7%
Instantaccess	Dialer	4.6%	3.4%
VB.AT	Worm	4.4%	3.4%
Fakerean	Rogue	4.1%	3.4%
Lolyda.AA1	PWS	2.3%	3.4%
C2LOP.gen!g	Trojan	2.1%	3.4%
Alueron.gen!J	Trojan	2.1%	3.4%
Lolyda.AA2	PWS	2.0%	3.4%
Dialplatform.B	Dialer	1.9%	3.4%
Dontovo.A	TojanDownloader	1.7%	3.4%
Lolyda.AT	PWS	1.7%	3.4%
Rbot!gen	Backdoor	1.7%	3.4%
C2LOP.P	Trojan	1.6%	3.4%
Obfuscator.AD	TojanDownloader	1.5%	3.4%
Malex.gen!J	Trojan	1.4%	3.4%
Swizzor.gen!I	TojanDownloader	1.4%	3.4%
Swizzor.gen!E	TojanDownloader	1.4%	3.4%
Lolyda.AA3	Dialer	1.3%	3.4%
Adialer.C	PWS	1.3%	3.4%
Agent.FYI	Backdoor	1.2%	3.4%
Autorun.K	Worm	1.1%	3.4%
Wintrim.BX	TojanDownloader	1.0%	3.4%
Skintrim.N	Trojan	0.9%	3.4%

When resizing the imagery data for the classification task, we consider the original image size in both datasets. For Big2015, where the original images were standardised to the shape of (128, 128, 3), we utilise this shape as well. For MALIMG, the original size varies, so we resize the data into a shape of (224, 224, 3), as this is the standard input size for the original ViT model.

4.2. Baselines and Evaluation Metrics. In our comparative analysis, we select three prominent CNN-based image classifiers to compare with our convolution-free approach. The first model, Inception [29], leverages parallel convolutional operations to effectively capture spatial hierarchies and

patterns, characterised by its impressive depth of 22 layers. The second model, MobileNet [30], is equipped with a novel efficient segmentation decoder, specifically designed for semantic segmentation, which delivers optimal performance even on mobile CPUs. The third model, Xception [31], employs a linear stack of depth-wise separable convolution layers with residual connections, which is followed by a point-wise convolution (1*1), implemented after the spatial convolution over each channel.

In the process of evaluation, we consider not only the overall accuracy but also the detection capability for minor classes and the equilibrium of performance. The metrics used are presented in Table 4, in which TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

Table 4. Basic evaluation metrics

Name	Equation
Recall/Detection Rate	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
AUC	
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
F1-score	$\frac{2*recall*precision}{recall+precision}$

Accuracy, a prevalent performance measure, compares the correctly predicted observations to the total observations. However, it can be misleading in imbalanced class distributions, particularly in security sectors such as malicious traffic detection and malware classification. The Area Under Curve (AUC) evaluates the performance across all conceivable classification thresholds. Precision, the ratio of correctly predicted positive observations to the total predicted positive observations, signifies a low false positive rate when high. Recall, on the other hand, is the ratio of correctly predicted positive observations to all actual positives, indicating sensitivity and detection ability. The F1 score, being the harmonic mean of precision and recall, serves as a better measure in cases of imbalanced classes. In multi-label classification tasks with imbalanced classes, such as our task of malware analysis, precision, recall, and F1 score are generally deemed more crucial metrics for evaluation.

4.3. Result Comparison. To evaluate the impact of the transformer block implemented in the classifier on the final performance, we test three variations with 4, 5, and 6 blocks implemented in the classifier, which are

represented by labels Trans-4, Trans-5, and Trans-6, respectively. Table 5 presents the overall performance on the test set of the Big2015 task. Generally, the convolution-free approaches outperform all the CNN models. Not only do our approaches have higher accuracy and AUC scores, but they also have relatively higher recall. While a model like Xception has a recall of only 0.3242, our approaches, Trans-4, Trans-5, and Trans-6, boast much higher recall scores. Furthermore, it is suggested that it may not be necessary to make the model extremely deep as the Trans-5, namely the model with 5 transformer blocks implemented, achieves the highest accuracy of 0.8369, AUC of 0.9791, and recall of 0.7959. Although the Trans-6 model has equally high accuracy, the highest precision of 0.8989, and F1 of 0.8359, the difference is not significant, indicating that the performance is not compromised by a lighter-weight model.

Table 5. Overall Performance on Test Set– Big2015

	Accuracy	AUC	Precision	Recall	F1
Inception	0.7592	0.9580	0.8480	0.6758	0.7522
MobileNet	0.7399	0.9470	0.8235	0.6563	0.7304
Xception	0.6774	0.9150	0.8384	0.3242	0.4676
Trans-4	0.8223	0.9698	0.8820	0.7666	0.8203
Trans-5	**0.8369**	**0.9791**	0.8792	**0.7959**	0.8355
Trans-6	**0.8369**	0.9777	**0.8989**	0.7813	**0.8359**

Conversely, the training process of our models could potentially be further improved. As presented in Figure 4, the trend of enhancing evaluation metrics stabilises during the fine-tuning phase; for example, our models consistently achieve lower loss rates. Specifically, Trans-5 exhibits the least loss and highest accuracy and AUC post the 46th epoch. Though our models showcase improved accuracy and reduced loss, it is evident that our model encounters notable variations throughout training, particularly in the validation loss curves. These fluctuations in loss optimisation may stem from the weight initialisation technique we employ. Nonetheless, it is worth noting that there is no sign of overfitting in our models. This indicates that despite the training fluctuations, our model remains dependable and robust, underscoring its efficient design and execution.

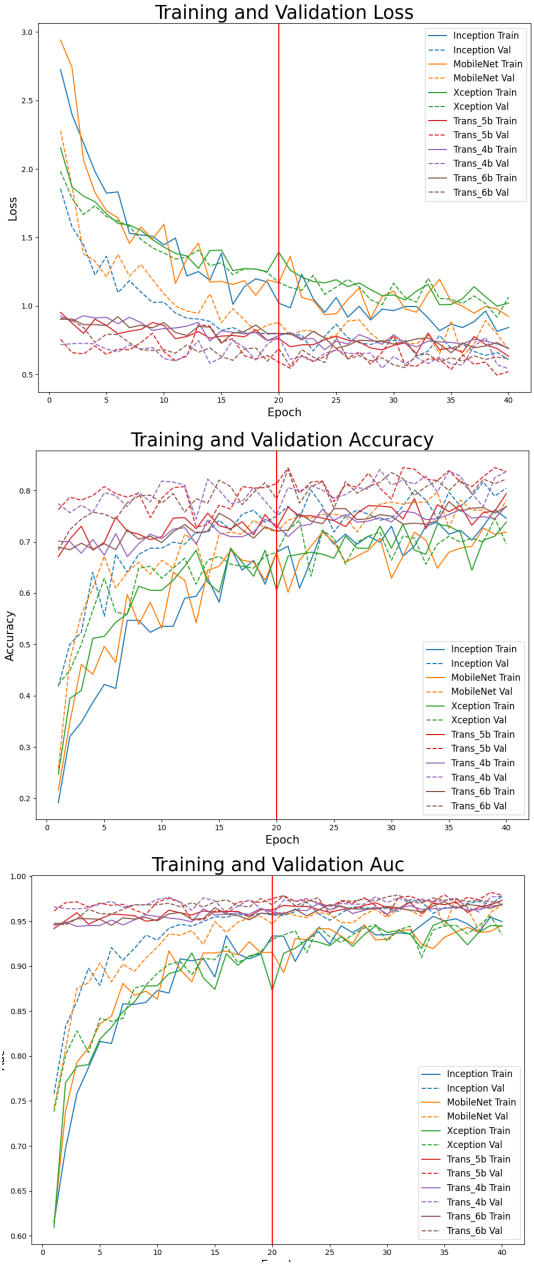


Fig. 4. The training process on the train and validation set of Big2015

Table 6 presents the outcomes of the test set for MALIMG, a particularly challenging task owing to its 25 categories. The findings reveal that all CNN models, along with Trans-4, slightly struggled with learning and accurate prediction, as indicated by their low accuracy and recall scores. In contrast, Trans-5 and Trans-6 exhibit superior performance. Although Trans-6, the deepest model, achieves the highest accuracy, recall, and F1 score, the discrepancies between Trans-5 and Trans-6 are not significant. With an accuracy of 0.9686, an AUC of 0.9992, a recall of 0.9686, and an F1 score of 0.9686, Trans-5 delivers satisfactory outcomes compared to the CNN models.

Table 6. Overall Performance on Test Set– MALIMG

	Accuracy	AUC	Precision	Recall	F1
Inception	0.8537	0.9011	0.8701	0.8418	0.8557
MobileNet	0.8537	0.9911	0.8701	0.8418	0.8557
Xception	0.8174	0.9008	0.8311	0.8311	0.8311
Trans-4	0.8581	0.9899	0.8601	0.8527	0.8564
Trans-5	0.9686	0.9992	0.9686	0.9686	0.9686
Trans-6	**0.9697**	**0.9998**	**0.9707**	**0.9697**	**0.9702**

Figure 5 presents the optimisation process for both the training and validation sets. In terms of loss and accuracy scores, all models exhibit continuous improvement during the initial training stage, although CNN models lag behind compared to Trans models. During the fine-tuning stage, the pronounced fluctuation of all curves highlights the increased challenge of training on tasks with more class labels. As MobileNet achieves the best training loss and accuracy but relatively poor validation loss and accuracy, there is a suggestion that this CNN-based model may be susceptible to overfitting. For Trans models, while the loss optimisation curve displays a consistent decrease, the accuracy experiences more significant fluctuations. This can be attributed to the weight initialisation process, where higher loss weights are assigned to minorities. Considering that accuracy evaluates overall performance, these fluctuations are to be expected. Nevertheless, a key point of optimism is that our model maintains an accuracy rate above 80% for the majority of the time, indicating its promising and reliable performance.

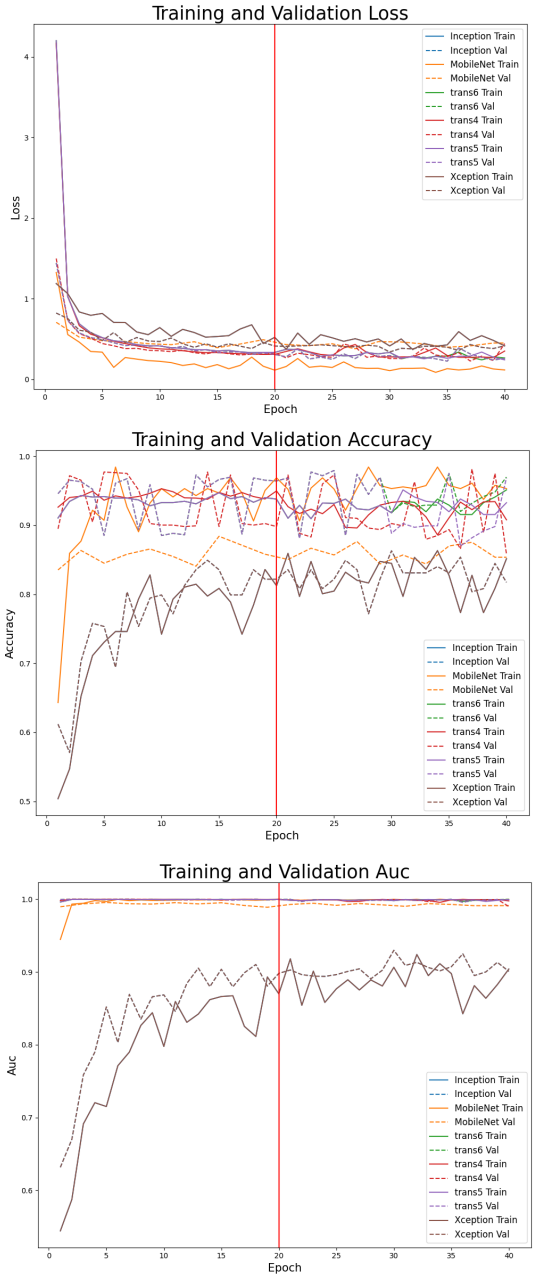


Fig. 5. The training process on the train and validation set of the MALIMG dataset

AUC presents a distinctive case. Except for Xception, all other models exhibit consistently high AUC scores across the entire training process. The reason behind the elevated AUC scores and the varying performance levels of the different models may be linked to the inherent nature of the AUC metric itself. While AUC evaluates the model's capacity to differentiate between positive and negative classes, the high AUC scores could possibly stem from the intrinsic characteristics of the data. Conversely, due to the imbalanced class distribution, the assessment of models' effectiveness through recall and the precision-recall equilibrium, known as the F1 score, can offer a more accurate evaluation of the performance.

CNN models exhibit satisfactory performance on the Big2015 task but fall short on the MALIMG task. In contrast, our approach consistently delivers superior results. This could be attributed to the capacity of self-attention-based models to manage long-range dependencies between pixels in images by assigning weights and prioritising patches for predictions. Consequently, these models can capture complex patterns and structures that may elude CNNs, particularly in intricate tasks requiring the prediction of 25 labels. Moreover, attention-based blocks are more readily parallelised, facilitating faster training times and enhanced performance. In scenarios involving extensive data, attention-based blocks also exhibit a less pronounced inductive bias. Nonetheless, Transformers may surpass CNNs in specific tasks or datasets. As highlighted in the original ViT paper, CNNs may still outperform transformers when dealing with smaller datasets.

4.4. Discussion on Future Research. Despite the demonstrated reliability and robustness of our model, there remain areas for potential enhancement. Primarily, the model's representation learning ability and detection capability could be further improved, as suggested by the results of the Big2015 task. Recent research has investigated various methods for transformer optimisation. For example, the Swin Transformer employs a shifted window-based self-attention mechanism, enabling it to capture both local and global dependencies in images [32]. Another notable study is DynamicViT [33], which introduces a technique that allows the model to adaptively adjust its computational complexity based on the complexity of the input image.

Another limitation is the insufficient consideration of obfuscation techniques, which can significantly impact the performance of malware classification systems. The technique employed for concealing malicious code makes accurate threat identification challenging. One possible enhancement involves augmenting the proposed method's resilience through feature engineering specifically tailored for obfuscation detection. In [34] the

authors undertook further static analysis to discern obfuscation patterns and devise discriminating features that distinguish benign applications from malware; this strategy enhanced the detector's efficacy in identifying obfuscated malware. An alternative approach is exploring adversarial ML techniques to simulate attacks on the classification model using obfuscated malware samples. Random noise can be generated and incorporated into image arrays to mimic obfuscation. This method can assist in identifying weaknesses within the classifier and subsequently refining the model to more effectively resist obfuscation techniques. Furthermore, extensive research into information-hiding techniques can be applied to image-based malware analysis. Several techniques exist for mitigating blurring to unveil hidden content or anomalies, including patch-line and fuzzy clustering-line priors for dehazing [35], as well as noise-aware filtering reversal through modified Landweber iterations [36].

Lastly, from the perspective of usable security, we intend to refine and expand upon our current methodology for malware classification. While the existing approach exhibits commendable performance in tasks involving multi-label classification, it encounters substantial challenges when confronted with novel or previously unidentified malware variants. Such instances are prone to misclassification or, more concerning, being erroneously identified as benign behaviours. Hence, our future endeavours will be directed towards the development and implementation of an advanced two-module framework. This innovative strategy will amalgamate a sophisticated multi-label classifier with an anomaly detection model. The latter component is planned to either embody an AE or leverage a probabilistic model specifically designed for outlier detection. In scenarios where the classifier is unable to accurately categorise a sample, defaulting to label it as benign, and concurrently, the anomaly detection module identifies it as an anomaly diverging from the established distribution patterns, such instances will be flagged as "potential new attacks". This designation will trigger alerts, thereby facilitating timely intervention and analysis. This strategic enhancement aims not only to bolster the accuracy and reliability of malware classification but also to establish a proactive defence mechanism against emerging cyber threats. By incorporating this two-pronged approach, our system will be better equipped to adapt to the evolving landscape of cybersecurity threats, ensuring enhanced protection for digital ecosystems.

5. Conclusion. In this study, we introduce a convolutional-free malware classifier, complemented by a GAN-based oversampler. This oversampling technique significantly amplifies the minority classes with realistic samples, while the classifier consistently surpasses CNN models

such as MobileNet, Inception, and Xception on a range of public datasets. Looking ahead, we aim to investigate the potential for enhancing the malware classification capability of our model and reducing its complexity. Nevertheless, the superior performance of our model across various tasks emphatically attests to its reliability and robustness. This not only substantiates our model's credibility but also positions it as a promising solution for tackling diverse and complex malware imagery classification tasks.

References

1. Altan G. SecureDeepNet-IoT: A deep learning application for invasion detection in industrial internet of things sensing systems. Transactions on Emerging Telecommunications Technologies. 2021. vol. 32. no. 4. DOI: 10.1002/ett.4228.
2. Tien C.-W., Chen S.-W., Ban T., Kuo S.-Y. Machine learning framework to analyze iot malware using elf and opcode features. Digital Threats: Research and Practice. 2020. vol. 1. no. 1. pp. 1–19. DOI: 10.1145/3378448.
3. Rizvi S., Aslam W., Shahzad M., Saleem S., Fraz M. Proud-mal: static analysis-based progressive framework for deep unsupervised malware classification of windows portable executable. Complex & Intelligent Systems. 2022. pp. 1–13.
4. Jung B., Kim T., Im E. Malware classification using byte sequence information. Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. 2018. pp. 143–148.
5. Alrawi O., Lever C., Valakuzhy K., Snow K., Monrose F., Antonakakis M., et al. The circle of life: A large-scale study of the IoT malware lifecycle. 30th USENIX Security Symposium (USENIX Security 21). 2021. pp. 3505–3522.
6. Smmarwar S., Gupta G., Kumar S. Android malware detection and identification frameworks by leveraging the machine and deep learning techniques: A comprehensive review. Telematics and Informatics Reports. 2024. vol. 14. DOI: 10.1016/j.teler.2024.100130.
7. Branitskiy A., Kotenko I. Network attack detection based on combination of neural, immune and neuro-fuzzy classifiers. IEEE 18th International Conference on Computational Science and Engineering. 2015. pp. 152–159.
8. Desnitsky V., Kotenko I., Nogin S. Detection of anomalies in data for monitoring of security components in the internet of things. XVIII International Conference on Soft Computing and Measurements. 2015. pp. 189–192.
9. Wang C., Zhao Z., Wang F., Li Q. A novel malware detection and family classification scheme for IoT based on deam and densenet. Security and Communication Networks. 2021. vol. 2021. no. 1. pp. 1–16. DOI: 10.1155/2021/6658842.
10. Yousefi-Azar M., Varadharajan V., Hamey L., Tupakula U. Autoencoder-based feature learning for cyber security applications. In 2017 International Joint Conference on Neural Networks (IJCNN). 2017. pp. 3854–3861.
11. Bakır H., Bakır R. Droidencoder: Malware detection using auto-encoder based feature extractor and machine learning algorithms. Computers and Electrical Engineering. 2023. vol. 110. DOI: 10.1016/j.compeleceng.2023.108804.
12. Venkatraman S., Alazab M., Vinayakumar R. A hybrid deep learning image-based analysis for effective malware detection. Journal of Information Security and Applications. 2019. vol. 47. pp. 377–389.
13. Yakura H., Shinozaki S., Nishimura R., Oyama Y., Sakuma J. Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017. pp. 55–56.

14. Li X., Wang L., Xin Y., Yang Y., Chen Y. Automated vulnerability detection in source code using minimum intermediate representation learning. Applied Sciences. 2020. vol. 10. no. 5. DOI: 10.3390/app10051692.
15. Wu P., Guo H., Buckland R. A transfer learning approach for network intrusion detection. IEEE 4th International Conference on Big Data Analytics. 2019. pp. 281–285.
16. Qiang Q., Cheng M., Zhou Y., Ding Y., Qi Z. Malup: A malware classification framework using convolutional neural network with deep unsupervised pre-training. In 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2021. pp. 627–634.
17. Hu J., Liu C., Cui Y. An improved cnn approach for network intrusion detection system. International Journal of Network Security. 2021. vol. 23. no. 4. pp. 569–575.
18. Xu Z., Fang X., Yang G. Malbert: A novel pre-training method for malware detection. Computers & Security. 2021. vol. 111(2). DOI: 10.1016/j.cose.2021.102458.
19. Habibi O., Chemmakha M., Lazaar M. Performance evaluation of cnn and pre-trained models for malware classification. Arabian Journal for Science and Engineering. 2023. vol. 48. no. 8. pp. 10355–10369.
20. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014. vol. 27. pp. 2672–2680.
21. Dong H., Kotenko I. Hybrid multi-task deep learning for improved iot network intrusion detection: Exploring different cnn structures. 2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS). 2024. pp. 7–12.
22. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser Ł., Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017. pp. 5998–6008.
23. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. arXiv preprint arXiv:2010.11929.
24. Huh M., Agrawal P., Efros A. What makes imagenet good for transfer learning? 2016. arXiv preprint arXiv:1608.08614.
25. Kingma D., Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
26. Loshchilov I., Hutter F. Decoupled weight decay regularization. 2017. arXiv preprint arXiv:1711.05101.
27. Ronen R., Radu M., Feuerstein C., Yom-Tov E., Ahmadi M. Microsoft malware classification challenge. 2018. arXiv preprint arXiv:1802.10135.
28. Nataraj L., Karthikeyan S., Jacob G., Manjunath B. Malware images: visualization and automatic classification. Proceedings of the 8th International Symposium on Visualization for Cyber Security. 2011. pp. 1–17.
29. Szegedy C., Ioffe S., Vanhoucke V., Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence. 2017. vol. 31. no. 1. DOI: 10.1609/aaai.v31i1.11231.
30. Howard A., Sandler M., Chen B., Wang W., Chen L.-C., Tan M., Chu G., Vasudevan V., Zhu Y., Pang R., Adam H., Le Q. Searching for mobilenetv3. IEEE/CVF International Conference on Computer Vision (ICCV). 2019. pp. 1314–1324.
31. Chollet F. Xception: Deep learning with depthwise separable convolutions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

32. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. pp. 10012–10022.
33. Rao Y., Zhao W., Liu B., Lu J., Zhou J., Hsieh C.-J. Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 13937–13949.
34. Kim D., Majlesi-Kupaei A., Roy J., Anand K., ElWazeer K., Buettner D., Barua R. DynODet: Detecting Dynamic Obfuscation in Malware. Proceedings of the Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA): 14th International Conference, DIMVA. 2017. pp. 97–118.
35. Liao M., Lu Y., Li X., Di S., Liang W., Chang V. An unsupervised image dehazing method using patch-line and fuzzy clustering-line priors. IEEE Transactions on Fuzzy Systems. 2024. vol. 4. pp. 1–15. DOI: 10.1109/TFUZZ.2024.3371944.
36. Wang L., Fayolle P., Belyaev A. Reverse image filtering with clean and noisy filters. Signal, Image and Video Processing. 2023. vol. 17. no. 2. pp. 333–341.

Huiyao Dong — Postgraduate, Faculty of information security, ITMO University; Programmer, laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: applied data science (particularly multi-task deep learning, autoencoder, imagery analysis, reinforcement deep learning), network security, IoT. The number of publications — 8. hydong@itmo.ru; 49 A, Kronverksky Av., 197101, St. Petersburg, Russia; office phone: +7(812)508-3311.

УДК 004.056 **DOI** 10.15622/ia.23.6.11

Х. Дун

КЛАССИФИКАЦИЯ ИЗОБРАЖЕНИЙ ВРЕДОНОСНЫХ ПРОГРАММ БЕЗ ИСПОЛЬЗОВАНИЯ СВЕРТОК С ИСПОЛЬЗОВАНИЕМ МЕХАНИЗМОВ ВНУТРЕННЕГО ВНИМАНИЯ

Дун Х. **Классификация изображений вредоносных программ без использования сверток с использованием механизмов внутреннего внимания.**

Аннотация. Анализ вредоносных программ является важнейшим аспектом кибербезопасности, направленным на выявление и дифференциацию вредоносного ПО от безвредных программ для защиты компьютерных систем от угроз безопасности. Несмотря на достижения в мерах кибербезопасности, вредоносные программы продолжают представлять значительные риски в киберпространстве, требуя точных и быстрых методов анализа. В этой статье представлен инновационный подход к классификации вредоносных программ с использованием анализа изображений, включающий три ключевых этапа: преобразование кодов операций в данные изображений RGB, использование генеративно-состязательной сети (GAN) для синтетической передискретизации и использование упрощенного классификатора на основе визуального трансформера (ViT) для анализа изображений. Данный метод повышает богатство функций и объяснимость с помощью данных визуальных изображений и устраняет несбалансированную классификацию с использованием методов передискретизации на основе GAN. Предложенная структура сочетает в себе преимущества сверточных автоэнкодеров, гибридных классификаторов и адаптированных моделей ViT для достижения баланса между точностью и вычислительной эффективностью. Как показали эксперименты, наш подход без использования сверток обладает превосходной точностью и прецизионностью по сравнению со сверточными моделями и превосходит модели CNN на двух наборах данных благодаря механизму многоголового внимания. На наборе данных Big2015 наша модель превосходит другие модели CNN с точностью 0,8369 и площадью под кривой (AUC) 0,9791. В частности, наша модель достигает точности 0,9697 и оценки F1 0,9702 на MALIMG, что является экстраординарным результатом.

Ключевые слова: обнаружение вредоносных программ, кибербезопасность, глубокое обучение, автоэнкодер.

Литература

1. Altan G. SecureDeepNet-IoT: A deep learning application for invasion detection in industrial internet of things sensing systems. Transactions on Emerging Telecommunications Technologies. 2021. vol. 32. no. 4. DOI: 10.1002/ett.4228.
2. Tien C.-W., Chen S.-W., Ban T., Kuo S.-Y. Machine learning framework to analyze iot malware using elf and opcode features. Digital Threats: Research and Practice. 2020. vol. 1. no. 1. pp. 1–19. DOI: 10.1145/3378448.
3. Rizvi S., Aslam W., Shahzad M., Saleem S., Fraz M. Proud-mal: static analysis-based progressive framework for deep unsupervised malware classification of windows portable executable. Complex & Intelligent Systems. 2022. pp. 1–13.
4. Jung B., Kim T., Im E. Malware classification using byte sequence information. Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems. 2018. pp. 143–148.

5. Alrawi O., Lever C., Valakuzhy K., Snow K., Monrose F., Antonakakis M., et al. The circle of life: A large-scale study of the IoT malware lifecycle. 30th USENIX Security Symposium (USENIX Security 21). 2021. pp. 3505–3522.
6. Smmarwar S., Gupta G., Kumar S. Android malware detection and identification frameworks by leveraging the machine and deep learning techniques: A comprehensive review. Telematics and Informatics Reports. 2024. vol. 14. DOI: 10.1016/j.teler.2024.100130.
7. Branitskiy A., Kotenko I. Network attack detection based on combination of neural, immune and neuro-fuzzy classifiers. IEEE 18th International Conference on Computational Science and Engineering. 2015. pp. 152–159.
8. Desnitsky V., Kotenko I., Nogin S. Detection of anomalies in data for monitoring of security components in the internet of things. XVIII International Conference on Soft Computing and Measurements. 2015. pp. 189–192.
9. Wang C., Zhao Z., Wang F., Li Q. A novel malware detection and family classification scheme for IoT based on deam and densenet. Security and Communication Networks. 2021. vol. 2021. no. 1. pp. 1–16. DOI: 10.1155/2021/6658842.
10. Yousefi-Azar M., Varadharajan V., Hamey L., Tupakula U. Autoencoder-based feature learning for cyber security applications. In 2017 International Joint Conference on Neural Networks (IJCNN). 2017. pp. 3854–3861.
11. Bakır H., Bakır R. Droidencoder: Malware detection using auto-encoder based feature extractor and machine learning algorithms. Computers and Electrical Engineering. 2023. vol. 110. DOI: 10.1016/j.compeleceng.2023.108804.
12. Venkatraman S., Alazab M., Vinayakumar R. A hybrid deep learning image-based analysis for effective malware detection. Journal of Information Security and Applications. 2019. vol. 47. pp. 377–389.
13. Yakura H., Shinozaki S., Nishimura R., Oyama Y., Sakuma J. Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017. pp. 55–56.
14. Li X., Wang L., Xin Y., Yang Y., Chen Y. Automated vulnerability detection in source code using minimum intermediate representation learning. Applied Sciences. 2020. vol. 10. no. 5. DOI: 10.3390/app10051692.
15. Wu P., Guo H., Buckland R. A transfer learning approach for network intrusion detection. IEEE 4th International Conference on Big Data Analytics. 2019. pp. 281–285.
16. Qiang Q., Cheng M., Zhou Y., Ding Y., Qi Z. Malup: A malware classification framework using convolutional neural network with deep unsupervised pre-training. In 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2021. pp. 627–634.
17. Hu J., Liu C., Cui Y. An improved cnn approach for network intrusion detection system. International Journal of Network Security. 2021. vol. 23. no. 4. pp. 569–575.
18. Xu Z., Fang X., Yang G. Malbert: A novel pre-training method for malware detection. Computers & Security. 2021. vol. 111(2). DOI: 10.1016/j.cose.2021.102458.
19. Habibi O., Chemmakha M., Lazaar M. Performance evaluation of cnn and pre-trained models for malware classification. Arabian Journal for Science and Engineering. 2023. vol. 48. no. 8. pp. 10355–10369.
20. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014. vol. 27. pp. 2672–2680.

21. Dong H., Kotenko I. Hybrid multi-task deep learning for improved iot network intrusion detection: Exploring different cnn structures. 2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS). 2024. pp. 7–12.
22. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017. pp. 5998–6008.
23. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. arXiv preprint arXiv:2010.11929.
24. Huh M., Agrawal P., Efros A. What makes imagenet good for transfer learning? 2016. arXiv preprint arXiv:1608.08614.
25. Kingma D., Ba J. Adam: A method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
26. Loshchilov I., Hutter F. Decoupled weight decay regularization. 2017. arXiv preprint arXiv:1711.05101.
27. Ronen R., Radu M., Feuerstein C., Yom-Tov E., Ahmadi M. Microsoft malware classification challenge. 2018. arXiv preprint arXiv:1802.10135.
28. Nataraj L., Karthikeyan S., Jacob G., Manjunath B. Malware images: visualization and automatic classification. Proceedings of the 8th International Symposium on Visualization for Cyber Security. 2011. pp. 1–17.
29. Szegedy C., Ioffe S., Vanhoucke V., Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence. 2017. vol. 31. no. 1. DOI: 10.1609/aaai.v31i1.11231.
30. Howard A., Sandler M., Chen B., Wang W., Chen L.-C., Tan M., Chu G., Vasudevan V., Zhu Y., Pang R., Adam H., Le Q. Searching for mobilenetv3. IEEE/CVF International Conference on Computer Vision (ICCV). 2019. pp. 1314–1324.
31. Chollet F. Xception: Deep learning with depthwise separable convolutions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.
32. Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. pp. 10012–10022.
33. Rao Y., Zhao W., Liu B., Lu J., Zhou J., Hsieh C.-J. Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 13937–13949.
34. Kim D., Majlesi-Kupaei A., Roy J., Anand K., ElWazeer K., Buettner D., Barua R. DynODet: Detecting Dynamic Obfuscation in Malware. Proceedings of the Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA): 14th International Conference, DIMVA. 2017. pp. 97–118.
35. Liao M., Lu Y., Li X., Di S., Liang W., Chang V. An unsupervised image dehazing method using patch-line and fuzzy clustering-line priors. IEEE Transactions on Fuzzy Systems. 2024. vol. 4. pp. 1–15. DOI: 10.1109/TFUZZ.2024.3371944.
36. Wang L., Fayolle P., Belyaev A. Reverse image filtering with clean and noisy filters. Signal, Image and Video Processing. 2023. vol. 17. no. 2. pp. 333–341.

Дун Хуэйяо — аспирантка, факультет информационной безопасности, Университет ИТМО; программист, лабораторией проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук. Область научных интересов: прикладная наука о данных (в частности, многозадачное глубокое обучение,

автокодирование, анализ изображений, глубокое обучение с подкреплением), сетевая безопасность, IoT. Число научных публикаций — 8. hydong@itmo.ru; Кронверкский проспект, 49 А, 197101, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

DOI 10.15622/ia.23.6.12

S. R. KRISHNAN, P. AMUDHA

ENHANCING VIDEO ANOMALY DETECTION WITH IMPROVED UNET AND CASCADE SLIDING WINDOW TECHNIQUE

R. Krishnan S., Amudha P. **Enhancing Video Anomaly Detection with Improved UNET and Cascade Sliding Window Technique.**

Abstract. Computer vision video anomaly detection still needs to be improved, especially when identifying images with unusual motions or objects. Current approaches mainly concentrate on reconstruction and prediction methods, and unsupervised video anomaly detection faces difficulties because there are not enough tagged abnormalities, which reduces accuracy. This paper presents a novel framework called the Improved UNET (I-UNET), designed to counteract overfitting by addressing the need for complex models that can extract subtle information from video anomalies. Video frame noise can be eliminated by preprocessing the frames with a Weiner filter. Moreover, the system uses Convolution Long Short-Term Memory (ConvLSTM) layers to smoothly integrate temporal and spatial data into its encoder and decoder portions, improving the accuracy of anomaly identification. The Cascade Sliding Window Technique (CSWT) is used post-processing to identify anomalous frames and generate anomaly scores. Compared to baseline approaches, experimental results on the UCF, UCSDped1, and UCSDped2 datasets demonstrate notable performance gains, with 99% accuracy, 90.8% Area Under Curve (AUC), and 10.9% Equal Error Rate (EER). This study provides a robust and accurate framework for video anomaly detection with the highest accuracy rate.

Keywords: anomaly detection, I-UNET, weiner filter, ConvLSTM, cascade sliding window, anomaly score.

1. Introduction. A significant task in video anomaly identification is recognising and localising unexpected occurrences in both place and time inside a video. These anomalies depict out-of-the-ordinary behaviours or events that may indicate possible concerns or security vulnerabilities. Depending on the context, anomalies may also be referred to as abnormalities, novelties, or outliers [1]. Unattended bags at airports, persons collapsing unexpectedly, or someone lingering suspiciously outside a guarded facility are some examples of video oddities [2]. Recognising and localising unusual occurrences in both place and time inside a video is a significant task in video anomaly identification. These anomalies describe unusual actions or events that may signal potential issues or security vulnerabilities [3]. Anomalies are also known as abnormalities, novelties, or outliers, depending on the context. This detection and analysis of anomalies is critical for improving security measures and addressing possible issues in various applications [4].

Anomaly detection in the video refers to automatically recognising aberrant events or behaviour within the spatiotemporal aspects of a video. It entails detecting actions or things that do not follow expected patterns or behaviours. Notably, the detection and localisation of video anomalies are

inextricably linked [5 – 6]. Real-time detection of anomalies in video data is critical because it allows immediate action to be taken upon recognising these anomalies, thereby preventing or mitigating adverse outcomes [7]. As a result, extensive research is being conducted to automate the process of detecting unusual occurrences in video surveillance systems. However, it might be hard to spot abnormalities in video broadcasts [8].

One important machine learning application is video anomaly detection, which looks for abnormal events or patterns in video data. This technology is critical in many areas, including security, surveillance, and industrial quality control [9]. It uses advanced algorithms and deep learning approaches to detect odd behaviours or events from the norm in a video frame. These systems can detect anomalies such as intruders in a secure facility, equipment breakdowns in manufacturing, or traffic accidents on the road by training models on massive datasets of usual events [10]. Machine learning has made significant advances in video anomaly detection. However, it still needs to improve, such as the necessity for substantial labelled data, the high processing intensity of deep learning techniques, and interpretability issues. These difficulties may make machine learning less accessible to smaller organisations or applications with limited resources [11]. Real-time processing requirements put existing infrastructure under pressure, and adaptation to changing video settings and anomalies can be constrained. As a result, while machine learning offers promise for video anomaly identification, careful analysis and resolution of these difficulties are required for successful implementation in varied applications [12].

Deep learning video anomaly detection is an advanced and sophisticated way of recognising odd events or behaviours inside video frames. This field of study uses the capabilities of deep neural networks, which are artificial intelligence systems designed to imitate the complicated functions of the human brain [13]. Deep learning for video anomaly detection entails training these neural networks to recognise and interpret typical activity patterns in video data. After learning what defines regular behaviour, the model can identify deviations from these established norms as anomalies or probable outliers [14]. Deep learning approaches, including CNNs and RNNs, are especially well-suited for detecting visual anomalies. CNNs thrive at analysing spatial information inside individual video frames, but RNNs excel at capturing temporal dependencies and event sequences. Combining these two types of neural networks allows the model to understand complicated spatiotemporal correlations, making it highly effective at detecting anomalies in video data [15].

Deep CNNs for video anomaly detection represent a cutting-edge way to detect unexpected events or anomalies inside video sequences. Deep CNNs have transformed Computer Vision tasks by automatically allowing models to learn and extract complicated spatial characteristics from pictures or video frames [16]. Deep CNNs excel in capturing intricate visual patterns and deviations when applied to video anomaly detection, making them a powerful tool in this field [17]. Video anomaly identification is a complicated task involving computer vision to identify unexpected motions or objects. Existing approaches concentrate on reconstruction and prediction but need help with obstacles such as low accuracy and complexity. A unique strategy for improving efficiency and accuracy by avoiding overfitting is suggested.

The following are the research work's key contributions:

1. The Improved UNET (I-UNET) is introduced in this work to improve anomaly detection in video frames by resolving overfitting and noise concerns, improving efficiency and accuracy.
2. Using a Weiner filter to preprocess video frames effectively reduces noise, resulting in cleaner frames for analysis and increased robustness for reliable anomaly identification.
3. The model employs an encoder-decoder architecture for efficient feature extraction, improved spatial and temporal information representation, and anomaly detection accuracy.
4. The Cascade Sliding Window Technique (CSWT) is utilised for anomaly detection in the post-processing phase, giving a sophisticated examination of frames and distinguishing between normal and abnormal ones.

The remaining manuscript is arranged as follows: The research strategy was explained in detail in the third segment, which also covers existing research. The fourth section simulates the suggested method and presents the research findings. Furthermore, a summary of the study's findings is given in the conclusion.

2. Literature Review. The research of video anomaly detection utilising deep learning and computer vision has experienced spectacular advances in recent years, with the emergence of various complicated algorithms. The researcher has provided several effective methods for detecting anomalies, as listed below.

The authors in [18] proposed a 3-stage ensemble-based unsupervised deep reinforcement algorithm for automated live video frames analytics, which employs a LSTM-based RNN for generating anomaly scores. The algorithm uses the least square method for optimal score creation, and model updates are accomplished by award-based reinforcement learning.

This method is intended for GPU and TPU-supported frameworks. However, the algorithm's capacity to handle many video frames simultaneously might be difficult, especially in high-demand applications, raising scalability difficulties.

A unique convolution autoencoder architecture for visual anomaly detection is presented in [19]. The architecture distinguishes between normalcy in appearance and motion behaviour and aberrant events by separating spatial and temporal information. The temporal autoencoder simulates optical flow using RGB difference, while the spatial autoencoder models normalcy by recreating the first individual frame. The method uses a deep Kmeans cluster strategy and a variance-based attention module to boost detection performance on rapidly moving outliers. However, limitations in accurately capturing complex motion patterns increase computational complexity, especially with quickly moving outliers.

Paper [20] frequently extracted low-level spatiotemporal features while ignoring semantic data. Deep learning algorithms, specifically CNN, are capable of extracting high-level information. A new hybrid visual embedding method was introduced for anomaly identification. The technique computes feature per frame with a pre-trained deep model, learns topic distributions with multilayer nonnegative matrix factorisation, and finds typical normal clusters with K-means. Experimental data demonstrate the method's usefulness in detecting anomalies. However, complex images with clutter or overlapping components and visual embedding may need help finding anomalies, possibly emphasising unimportant parts or failing to detect minor ones.

A residual spatiotemporal autoencoder for anomaly identification in security footage is suggested by the [21]. The method takes advantage of normalcy modelling to find departures from standard patterns. The trainable end-to-end autoencoder uses reconstruction loss to detect aberrant frames. Regarding cross-dataset generalisation, residual blocks work better than deeper layers because of their incremental effectiveness. However, the proposed technique has limitations such as adaptability, complexity, and generalizability to diverse datasets compared to deeper layers.

In study [22] the authors introduced an EADN deep learning-based approach. It splits video into prominent shots, extracts spatiotemporal information with a CNN, and learns spatiotemporal features with LSTM cells. The model's utility is demonstrated by extensive testing on benchmark datasets and comparisons with the most advanced techniques. Nonetheless, there is still room in the EADN design for improving efficiency and accuracy in real-time.

In paper [23] created a video anomaly detection system that depends on unsupervised frame prediction and enhances overall performance. The method is based on a U-Net-like structure that consists of a memory module for storing standard patterns, a Time-distributed 2D CNN-based encoder and decoder, and a multi-branch structure for extracting contextual information. However, this method may result in overfitting.

In paper [24] the authors projected a reliance on the reconstruction or prediction of future frames. In most approaches, accuracy is impacted by the requirement for more excellent temporal continuity between video frames. Using a hybrid dilated convolution module and DB-ConvLSTM module, a novel technique combines these two models. Experiments show this approach detects abnormalities more correctly in diverse video settings than state-of-the-art technologies. However, the completeness of the training data for each scenario is a prerequisite for the proposed model.

Paper [25] suggested a deep CNN encoder & multi-stage channel attention decoder for autonomous anomaly detection in video surveillance systems. Temporal shift methods and channel attention modules were used for contextual dependency extraction. However, the proposed method is computationally expensive, especially for high-resolution videos.

In paper [26] a novel anomaly detection approach for surveillance operations, focused on mobile cameras. Three techniques were employed to extract robust features from Unmanned Aerial Vehicle (UAV) footage: One-Class Support Vector Machine (OCSVM), two manually constructed approaches called Histogram Oriented Gradient (HOG) and Histogram Oriented Gradient 3 Dimensional (HOG3D), and a pre-trained CNN.

The model by the author in [27] for detecting anomalies was ineffective because of significant differences within and across classes. Two novel multi-view representation learning approaches were proposed: a hybrid multi-view representation learning that combined robust handcrafted features with deep features from 3D-STAE and a deep multi-view representation learning that combined features from two-frames SpatioTemporal AutoEncoder and deep features from 3D-STAE. The video anomaly detection with several existing works is tabulated in Table 1.

As a result, managing several video frames at once may present difficulties for the anomaly detection method in complicated image processing, particularly in high-demand applications. Its ability to precisely capture intricate motion patterns and growing computational complexity is likewise limited. Notwithstanding these drawbacks, the EADN design can be strengthened to increase real-time accuracy and efficiency.

Table 1. Video Anomaly Detection with state-of-art-of techniques

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[18]	Deep Kmeans cluster strategy	Convolutional autoencoder for anomaly detection in videos, separating spatial and temporal information to capture appearance and motion behaviour separately	Enhances anomaly detection accuracy by effectively capturing appearance and motion behaviour independently by dissociating spatial and temporal representations	It may introduce additional computational complexity, potentially requiring more resources for training and inference	9% reduction in EER of UCSD Ped1, a 13% reduction in ERR of UCSD Ped2 and a 4% improvement in accuracy in both datasets
[19]	CNN	Aims to develop a CNN-based methodology for automatically detecting traffic accidents in surveillance videos from video traffic surveillance systems	Automated detection of traffic accidents in surveillance videos, reducing reliance on manual monitoring and enabling prompt emergency response	It potentially limits its performance in detecting accidents that must be adequately represented in the training data	AUC (UCSD Ped2 dataset): 96.7% AUC (Avenue dataset): 87.1% AUC (ShanghaiTech): 73.7%
[20]	CNN	Develop a residual spatiotemporal autoencoder for anomaly detection in surveillance videos, leveraging normality modelling to identify irregularities as deviations from standard patterns	The method uses normality modelling and reconstruction loss to identify abnormal spatiotemporal events in surveillance videos accurately	It can be challenging in complex and dynamic surveillance environments, potentially leading to false positives or negatives	UCSD Ped1 dataset: EER=8.1 AUC=93.9 Accuracy=90.3 UCSD Ped2 dataset: EER=6.1 AUC=97.3 Accuracy=95.4
[21]	Residual spatiotemporal autoencoder	To enhance anomaly detection by integrating reconstruction and future frame prediction models, addressing limitations in existing methods	Captures spatial features at various scales, enabling the model to effectively detect anomalies by considering objects of different sizes and complexities in surveillance videos	The increased model complexity could lead to overfitting	AUC (Avenue dataset): 0.82 AUC (LV dataset): 0.63

Continuation of the Table 1

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[22]	Deep convolutional neural network-based encoder and a multi-stage channel attention-based decoder	Aims to develop an advanced anomaly detection system for video surveillance by effectively integrating spatial and temporal information	The proposed method effectively captures spatial and temporal features, enhancing anomaly detection accuracy in surveillance videos	It may lead to increased computational complexity, requiring significant resources for training and inference	UCSDped1 Accuracy:93 False alarm rate: 0.08 UCSDped2 Accuracy:97.0 False alarm rate:0.06 CUHK Avenue dataset Accuracy:97.0 False alarm rate: 0.04 UCF-Crime dataset Accuracy:98.0 False alarm rate:0.03
[23]	Convolutional Neural Network (CNN) and two popular handcrafted methods (Histogram of Oriented Gradient (HOG) and HOG3D). One Class Support Vector Machine (OCSVM)	To address the limitations of existing stationary camera surveillance systems in anomaly detection by proposing new techniques suitable for Unmanned Aerial Vehicle (UAV)-based surveillance missions	Enhancing anomaly detection by capturing comprehensive surveillance footage from various angles and viewpoints	Utilising multiple feature extraction methods increases computational complexity and resource requirements	UCSDped1 AUC:83.8 EER:22.2 UCSDped2 AUC:97.6 EER:6.6 Avenue dataset AUC:89.0 EER:18.1
[24]	3D spatiotemporal autoencoder	To enhance automatic surveillance of human activities by addressing the challenges posed by complex real-time scenarios, such as camera movements, cluttered backgrounds, and occlusion	The method captures high-level semantic information and fine-grained details, providing a more comprehensive representation of surveillance video data	increases the computational complexity and resource requirements of the proposed methods	AUC raises 2.0%, 1.2%, and 1.6% for UCSD Ped1, UCSD Ped2, and CUHK Avenue datasets compared with it
[25]	attention-based residual autoencoder	It efficiently utilises spatial and temporal information by adopting both spatial and temporal branches in a single network	The model exceeds the state-of-the-art results on three standard benchmark datasets, even without an optical flow detector	It may be generalised to 3D data for real-world engineering applications	This model achieved 97.4% for UCSD Ped2, 86.7% for CUHK Avenue, and 73.6% for the ShanghaiTech dataset in terms of AUC

Continuation of the Table 1

No.	Technique	Objectives	Advantages	Limitation/ future scope	Result
[26]	OCSVM, HOG, HOG3D, CNN	The goal is when a mobile camera records videos for a surveillance mission with the assistance of a UAV	The potential of UAVs to provide an original aerial perspective is one of its primary advantages	The upper layers of the pre-trained CNN can also be adjusted using transfer learning to better match the target problem in future	HOG model Recall:1 Precision:0.7070 F1 score:0.8284 Accuracy:78.97 PCA-HOG model Recall:1 Precision:0.7073 F1 score:0.8286 Accuracy:79.00 HOG3D model Recall:1 Precision:0.8421 F1 score:0.9143 Accuracy:90.13 GoogleNet model Recall:1 Precision:0.8837 F1 score:0.9383 Accuracy:93.57
[27]	Hybrid multi-view representati on learning	This model uses handcrafted spatiotemporal autocorrelation of gradient, and raw video segments are taken as input for learning the regular patterns in surveillance videos	This model can be accomplished by extracting features from various representations (views) of the raw input data	In the future, anomalies' dependency will be captured in context, leading to better discrimination and overall performance	Avenue dataset Accuracy:82.4 AUC:0.83 LV dataset: Accuracy:64.9 AUC:0.60 BEHAVE dataset Accuracy:80.05 AUC:0.81

Nonetheless, overfitting could happen, and the model depends heavily on the completeness of the training set. The suggested approach is computationally costly, mainly when used in high-resolution videos.

3. Proposed work. Video anomaly detection and segmentation in smart cities is a crucial computer vision problem for smart surveillance and public safety. However, existing research faces challenges such as limited scalability, difficulty detecting complex anomalies, and potential overfitting. The model is computationally demanding and relies on training data accuracy, leading to low efficiency, accuracy, and overfitting in identifying and segmenting video anomalies

The proposed approach overcomes the problem of overfitting in anomaly detection within video frames by introducing a novel I-UNET. Overfitting happens when a model becomes too specialised to the training data, resulting in decreased efficiency and accuracy when applied to fresh,

previously unseen data. The I-UNET is meant to detect irregularities in video frames, which improves detection accuracy. Video frames frequently contain a significant amount of noise, which can impede the accurate detection of anomalies. To overcome this issue, a Weiner filter is applied to the frames during preprocessing to reduce noise. This step seeks to improve the input data quality and the overall presentation of the anomaly detection technique. During anomaly detection, the proposed approach considers spatial and temporal information. To accomplish this, a Convolutional Long Short-Term Memory (ConvLSTM) is added to the model. The ConvLSTM allows the model to consider both the spatial properties of the video frames and the temporal dependencies between consecutive frames, resulting in a more complete comprehension of the information. The suggested model uses the cascade sliding window technique (CSWT) to produce an anomaly score during the post-processing stage. The CSWT analyses the video frames and assigns a score reflecting the chance of an abnormality. This anomaly score was utilised to determine whether a specific frame contains an anomaly or is within the normal range. The suggested diagram's overall architecture is depicted in Figure 1 below.

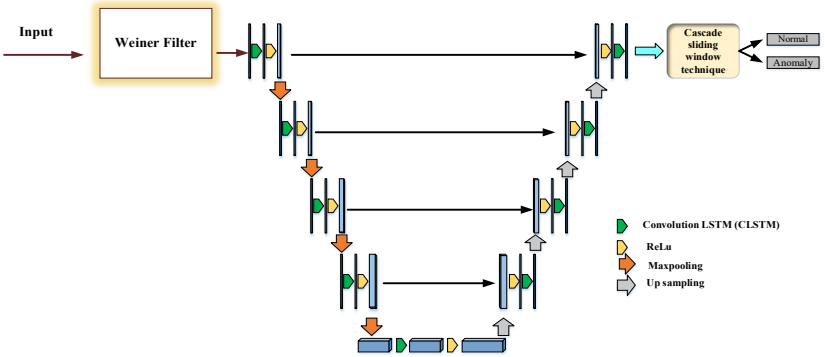


Fig. 1. Architecture diagram of the proposed model

3.1. Weiner Filter for pre-processing. The Weiner filter is used to reduce noise from images [28]. Weiner filters use Linear Time-Invariant (LTI) filtering of an observed noisy process with known stationary signal, noise spectra, and additive noise to estimate a desired or target random process. The Weiner filter decreases the mean square error between the estimated and calculated random processes. Using a related signal as an input and filtering it to get the approximation as an output, the Weiner filter computes a statistical estimate of an unknown signal. Stated differently, the Weiner filter is an adaptive filter that determines the neighbourhood's mean

and variance before applying a lower level of smoothing when the variation is significant and a higher level of smoothing when it is negligible.

The filter reduces the error between the expected and original signals. The error measure, given an original image and a processed image in Figure 7, represents the original and pre-processed images with a wiener filter.

The filter lessens the variation between the original and approximated signals. The error measure for an original image f and a processed image $\hat{f}$ is as follows in equation (1):

$$e^2 = E\{(f - \hat{f})^2\}, \tag{1}$$

where E{.} is the argument's predictable value, generating an approximated image boils down to locating the quadratic error function's minimum. The frequency domain is used to accomplish this, and the following presumptions are made: the image and noise have a zero mean, the noise and image are uncorrelated, and a linear function reduces the intensity levels in the expected picture. Depending on these circumstances, the error function's minimum is provided in equation (2):

$$\hat{F}(u,v) = \left[\frac{H^*(u,v)S_f(u,v)}{S_f(u,v)|H(u,v)|^2 + S_f(u,v)}\right] G(u,v), \tag{2}$$

where $\hat{F}(u,v)$ represent the predictable image in the frequency domain, $H(u,v)$ denote the transform of the degradation function, $G(u,v)$ denote the transform of the degraded image, $H^*(u,v)$ denote the complex conjugate of $H(\mathrm{u},\mathrm{v})$ and $S_f(u,v)$=$|\mathrm{F}(\mathrm{u},\mathrm{v})|^2$ is the power spectrum of the non-degraded image. The magnitude of the complex value squared represents the result of multiplying a complex value by its conjugate, according to the filter's general principle. Consequently, in equation (3):

$$\hat{F}(u,v) = \left[\frac{1}{H(u,v)} \frac{|H(u,v)|^2}{|H(u,v)|^2 + S_\eta(u,v)/S_f(u,v)}\right] \quad G(u,v), \tag{3}$$

where $S_\eta(\mathrm{u},\mathrm{v}) = |N(u,v)|^2$ represent the power spectrum of noise. The term $S_\eta(u,v)/S_f(u,v)$ is substituted by a constant K due to the rarity of knowing the non-degraded image's power spectrum.

The Weiner filter can correct digital image processing noise caused by continuous power additive noise. Thus, the neighbourhood size and noise power are the parameters of the Weiner filter.

Figure 2 depicts both the original and pre-processed images. The Weiner filter method is used in pre-processing to eliminate noise from the frame.

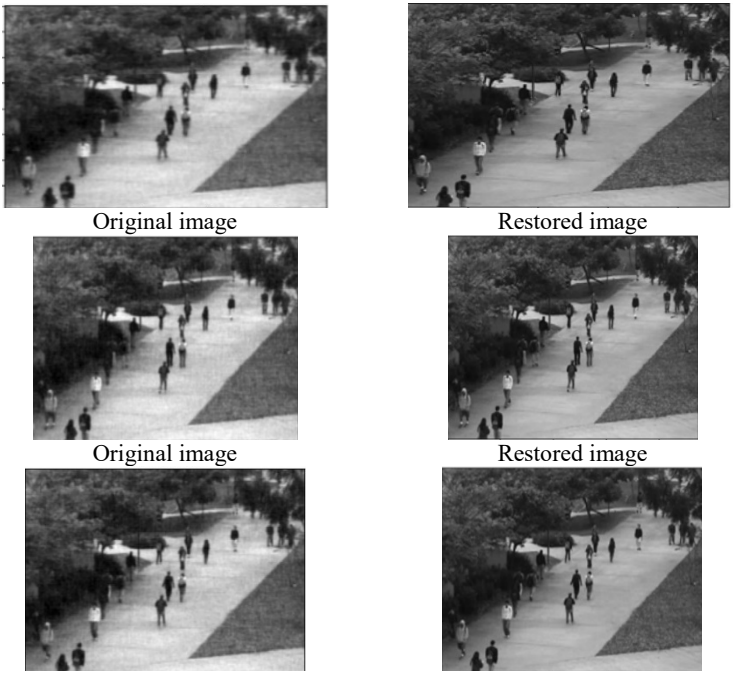


Fig. 2. Original and Pre-processed images with Weiner filter

3.2. Improved UNET for feature segmentation

3.2.1. U-NET. The U-Net architecture is a popular and useful paradigm for segmenting video images. It is U-shaped, symmetrically decoder-path and encoder-path [29]. Because of its U-shaped design, the model can record local information and information about the larger surroundings. In the encoder approach, convolutional and pooling layers are utilised to gradually downscale the input image, which aids in extracting high-level features and collecting contextual data. Each down-sampling step reduces the spatial size of the feature maps while improving their depth. Information is transferred from the encoder to the decoder using skip connections. These links connect the essential layers between the encoder and decoder routes. The skip connections allow the decoder to access high-resolution information from the encoder while acting as a gradient flow

shortcut during training. The decoder path performs feature map up-sampling using deconvolutional layers or up-sampling followed by convolutional layers. This method retains the contextual information the encoder learned as the feature maps' spatial resolution is progressively restored. Skip connections enhance segmentation results by merging feature maps from the encoder and decoder.

3.2.2. I-UNET. When renovating from standard convolution layers to ConvLSTM [30] layers inside the U-NET structure, it is vital to account for the inherent spatiotemporal dependencies in video data. In this improved approach, the Encoder with ConvLSTM is used strategically to capture spatial dependencies over the entire video frame. Concurrently, the decoder is upgraded with ConvLSTM layers to manage temporal dependencies during decoding properly. The introduction of ConvLSTM layers allows the model to include spatial properties within individual frames and temporal correlations between subsequent frames. As a result, the last layer of the decoder is modified to provide an anomaly prediction map for each frame. This holistic approach enables the U-Net to more comprehensively understand and leverage the spatiotemporal intricacies inherent in video data, making it well-suited for tasks such as video anomaly detection. The architecture of the Improved U-NET is displayed in Figure 3.

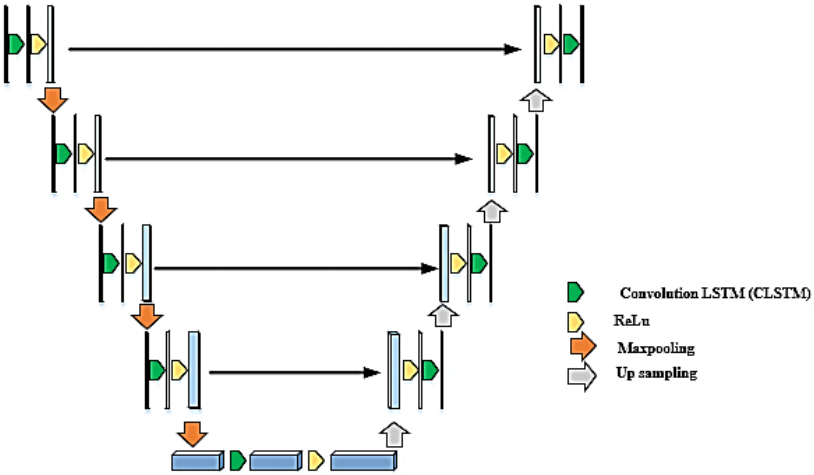


Fig. 3. Improved U-NET Architecture

3.2.2.1. ConvLSTM. Convolutional Long Short-Term Memory, or "ConvLSTM" architecture, combines the best features of Long Short-Term Memory (LSTM) networks with the concepts of Convolutional Neural

Networks (CNNs) to improve processing speed for sequential data, such as video frames. The forget gate is an essential component of classic LSTM networks that helps choose, keep, or discard data from earlier steps. By addressing the vanishing gradient issue, this method helps the network to handle longer sequences and aggregate higher-level information efficiently. Enhancing this concept, the ConvLSTM model substitutes convolutions for conventional matrix operations. The creation of spatial feature maps is enhanced by this adjustment, which lowers the model's parameter count (weights). The ConvLSTM model is especially well-suited for tasks like video frame prediction because it works with convolutions to capture spatial dependencies within the input data effectively. Sequence modelling for video data is greatly advanced by the convolutional procedures combined with LSTM units. The model's capacity to recognise intricate temporal patterns and spatial correlations in video data is improved by this method, which produces a more efficient and spatially aware representation of the input sequences. Convolutions are used instead of matrix operations in ConvLSTM instead of the typical fully connected LSTM (FC-LSTM). ConvLSTM uses convolution for hidden-to-hidden and input-to-hidden connections, which results in better spatial feature maps with less weight requirements. The following equations (4-9), which explain how data moves through the network and how the forget gate, input gate, and output gate are calculated, can be used to sum up the working principles of the ConvLSTM unit:

$$f_t = \sigma(W_f * [h_{t-1}, x_t, C_{t-1}] + b_f, \quad (4)$$

where f_t indicates the forget gate activation vector at time step t, σ is the sigmoid function, the input vector is represented by the variable x_t, the hidden state is represented by h_t, b_f represents the bias vector for the forget gate, W_f represents the weight matrix for the forget gate and the cell state is represented by C_t at times *t*.

$$i_t = \sigma(W_i * [h_{t-1}, x_t, C_{t-1}] + b_i, \quad (5)$$

where i_t indicates the Input gate activation vector at time step t, W_i represents the weight matrix for the input gate, b_i denotes the bias vector for the input gate.

$$\hat{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C, \quad (6)$$

where $\hat{C}_t$ represents the candidate cell state at time step t, the hyperbolic tangent function is represented by $\tanh$, W_C denotes the weight matrix for the candidate cell state, the bias vector for the candidate cell state is denoted by b_C.

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \hat{C}_t, \tag{7}$$

where C_t is the new cell state at time step t, f_t represents the forget gate activation vector at time step t, the previous cell state at time step $t-1$ is indicated by C_{t-1}, i_t denotes the input gate activation vector at time step t. The sign $\otimes$ indicates the Hadamard product.

$$o_t = \sigma(W_o * [h_{t-1}, x_t, C_{t-1}] + b_0), \tag{8}$$

where o_t is the output gate activation vector at time step t, the weight matrix for the output gate is represented by W_o, b_0 is the bias vector for the output gate.

$$h_t = o_t \otimes tanh(C_t), \tag{9}$$

where h_t is the hidden state at time step t, the output gate activation vector at time step t is indicated by o_t, $tanh$ represents the hyperbolic tangent function, C_t is the cell state at time step t.

Convolutional filters replace the set of weights for each link in the input (the symbol * denotes a convolution operation). Because of its capacity to convey spatial attributes temporally through each ConvLSTM state, ConvLSTM works better with images than fully connected LSTM.

3.3. Cascade Sliding Window Technique for classification. Cascade sliding window is a technique used in object detection that can also be applied to video anomaly detection. Finding strange occurrences or behaviours in video frames is the goal of video anomaly detection. The cascade sliding window technique employs a multi-stage process to look for objects or abnormalities inside each frame at various scales and places.

The cascade sliding window method is a methodical way to get an anomaly score in the context of video analysis. This method is intended to determine whether or not a specific frame in a video sequence contains an anomaly. The process entails scanning each frame at different positions and scales using a sliding window. A classifier is used at each stage of this procedure to evaluate the information within the sliding window, discriminating between typical and abnormal patterns. The cascade

structure incorporates numerous layers of classifiers, each contributing to determining the anomaly score. The first stage usually uses simpler and faster classifiers to quickly exclude non-anomalous regions, decreasing the computational cost for the following stages. More advanced classifiers are used to refine the anomaly detection as the process progresses along the cascade. This methodology's anomaly score is a critical output as a quantitative measure to infer the likelihood of an anomaly's presence in a given frame. By defining appropriate criteria for anomaly scores, it is feasible to make educated decisions about whether a frame exhibits anomalous behaviour. This allows for effective identification and analysis of odd events in the video frame. The cascade sliding window framework contributes to accuracy and processing efficiency in the anomaly identification process, making this method an organised and efficient way of distinguishing anomalies in video data. In Algorithm 1, the Cascade Sliding Window technique is illustrated.

The frame size R represents a frame's height and width in Algorithm 1, and the window size is indicated by $\hat{R}$. The technique generates an image I by squaring the distinction between an actual frame and one that has been predicted. This method is chosen because it allocates higher values to pixels in the abnormal region of I rather than making use of the fundamental distinction between an expected and actual frame. Compared to the distinction between a genuine frame and a prediction frame, it is more successful at identifying aberrant frames. The average of an anomalous frame with a small abnormal section may resemble or be less than that of a regular frame if the anomalous frame employs the difference between an actual frame and a forecasted frame. The window on $\boldsymbol{I}$ starts to move to the right as much as $\widehat{\boldsymbol{R}}$ at positions x=0 and y=0. The window travels to the left side of I and up to $\widehat{\boldsymbol{R}}$ if it reaches the right side of $\boldsymbol{I}$. The window will then start to slide to the right by the same amount as $\widehat{\boldsymbol{R}}$ as you lower $\widehat{\boldsymbol{R}}$ as the window size v decreases. Continue in the same manner until the window reaches $\boldsymbol{I}$'s upper right corner. In case the window is unable to get either the top or right sides of $\boldsymbol{I}$ due to the remaining space in $\boldsymbol{I}$ being less than the window, it will go to $\boldsymbol{y = R - \widehat{R}}$ for the top side of $\boldsymbol{I}$ and to $\boldsymbol{x = R - \widehat{R}}$ for the right side $\boldsymbol{I}$. It is provided in lines 9 and 18 of Algorithm 1. Determine the average for a frame $\boldsymbol{P_k}$ that corresponds to the moving window. It is comparable to $\boldsymbol{P_k}$, the mean squared error (MSE) between a real frame and a prediction frame. When the window reaches the upper right corner of $\boldsymbol{I}$, take n frames from the front of $\boldsymbol{P_k}$ and sort $\boldsymbol{P_k}$ in increasing order. The anomalous score S is then calculated by averaging the n patches.

Algorithm 1. Cascade Sliding Window

Input: actual present frame $F_{i,j,}$ anticipated present frame $\hat{F}_{i,j,}$ frame size R, window size $\hat{R}$, window decrease size v

Results: score S for anomalies

1 Set up the coordinates x=0 and y=0 and the P_k mean for each frame.

2 $I_{i,j} = (F_{i,j} - \hat{F}_{i,j})^2$ /* image on the square of the difference between $F_{i,j}$ and $\hat{F}_{i,j}$ */

3 While y < R do

4 if $y + \hat{R} \leq R$ then

5 while $x < R$ do

6 if $x + \hat{R} \leq R$ then

7 $P_k = \frac{1}{\hat{R}^2} \sum_{i=x}^{x+\hat{R}} \sum_{j=y}^{y+\hat{R}} I_{i,j}$

8 else

9 $P_k = \frac{1}{\hat{R}^2} \sum_{i=R-\hat{R}}^{R} \sum_{j=y}^{y+\hat{R}} I_{i,j}$

10 end if

11 $x = x + \hat{R}$

12 end while

13 else

14 while x < R do

15 if $x + \hat{R} \leq R$ then

16 $P_k = \frac{1}{\hat{R}^2} \sum_{i=x}^{x+\hat{R}} \sum_{j=R-\hat{R}}^{s} I_{i,j}$

17 else

18 $P_k = \frac{1}{\hat{R}^2} \sum_{i=R-\hat{R}}^{R} \sum_{j=R-\hat{R}}^{R} I_{i,j}$

19 end if

20 $x = x + \hat{R}$

21 end while

22 end if

23 $x = 0$

24 $y = y + \hat{R}$

25 $\hat{R} = \hat{R} - v$

26 end while

27 arrange (P_k) in ascending order

28 $S = \frac{1}{n} \sum_{i=1}^{n} P_i$

29 Return S

The cascade sliding window's decreasing window size, shown by v in line 25 of Algorithm 1, is essential. It assumes that objects get smaller and farther away from the items under video monitoring. Contrasting it with another approach that uses the MSE between an actual and a forecasted present frame shows how well the cascade sliding window technique performs.

3.3.1. Anomaly detection. Using the cascade sliding window, compute the anomaly score for every frame. The anomaly score of the I-UNET model output frame runs from 0 to $colordepth^2$, and the colour depth in the framework is 256. Consequently, the range of the anomaly score generated by the cascade sliding window is 0 to 65536. This range is too broad to establish the anomalous frame threshold. As a result, the anomaly score must be normalised. The anomaly score between 0 and 1 is normalised using the following formulas equation (10):

$$S'(t) = 1 - \frac{S(t) - min_t S(t)}{max_t S(t) - min_t S(t)}, \quad (10)$$

S'(t) is the normalised anomaly score, and S(t) is the anomaly score for frame t. Videos have two anomaly scores: $min_t S(t)$ and $max_t S(t)$ for maximum and minimum anomaly scores, respectively. However, while obtaining a new frame in the actual world, the $max_t S(t)$ and $min_t S(t)$ values could change. It results in a recalculation of the threshold and the acquired anomaly scores. Normalising the anomaly score from 0 to 1 will solve this problem using equation (11).

$$S'(t) = \frac{S(t)}{colordepth^2}, \quad (11)$$

where colour depth refers to the colour depth of the R-Net model's output frame, even if the maximum and minimum anomaly scores are altered, this normalisation does not necessitate recalculating the threshold and anomaly scores.

4. Results and Discussions. The model was trained using thousands of video frames from a video dataset, and a thorough testing procedure was conducted to ascertain the efficacy of the novel approach.

4.1. Datasets. UCSD Ped2. A stationary camera positioned at a height that provided a view of pedestrian routes was utilised to collect the UCSD anomaly detection dataset. There was a range in the walkways' population density from sparse to congested. In its original state, the video only features pedestrians. Either non-pedestrian objects moving through the walkways or abnormal pedestrian movement patterns cause abnormal events. Individuals, skateboarders, bicyclists, and small carts are frequently observed strolling down a path or in the adjacent grass. There were also a few reported instances of wheelchair-using individuals. Since none of the anomalies were created to compile the dataset, they are all-natural. Two subgroups were created from the data, each representing a distinct scene. Each sequence's video clip was segmented into segments of

about 200 frames. Scenes with pedestrian movement parallel to the camera plane are categorised as Peds2, which includes twelve testing and sixteen training video examples. The ground truth annotation for every clip contains a binary flag for every frame that indicates whether an abnormality is present in that particular frame. Furthermore, pixel-level binary masks, including anomaly zones, are manually created for a subset of 12 clips for Peds2. This is meant to make it possible to assess how well algorithms perform in terms of their capacity to localise anomalies. Here, the data is split into 60% for training and 40 % for testing. (https://www.kaggle.com/datasets/karthiknm1/ucsd-anomaly-detection-dataset).

4.2. Performance Evaluation. This segment demonstrates how the suggested I-UNET approach may successfully classify video anomaly image frames designated as normal or anomalous. The suggested model's accuracy and loss analyses are displayed during the training process across the 8th epochs. The proposed model improves accuracy while losing utility. It demonstrates that the proposed model converges very quickly.

The I-UNET method trains a model across eight epochs with the Adam optimizer, consistently obtaining 99% accuracy, as shown in Figure 4. This high degree of accuracy shows how reliable and efficient the I-UNET technique is in correctly identifying and analysing data patterns. The model's high accuracy indicates its potential for various applications where precision is crucial.

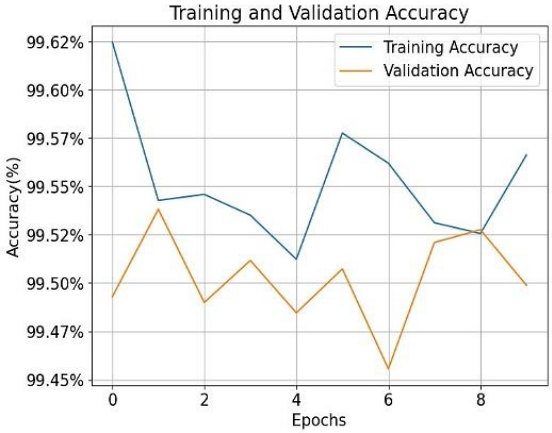


Fig. 4. Training and validation Accuracy

Employing the Adam optimizer, the training and validation AUC of the I-UNET technique was monitored throughout eight epochs. With an astounding 90.8% AUC, the model successfully identified normal and anomalous instances in the video data. This high AUC score demonstrated the strength and adaptability of the technique over all epochs, as shown in Figure 5. The Adam optimiser's practical training probably aided the model's convergence towards an optimal solution, highlighting the dependability and capacity to generalise the model and increasing the likelihood of precise anomaly discovery.

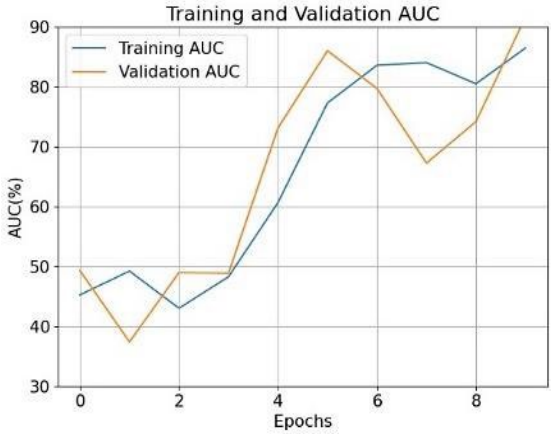


Fig. 5. Training and validation AUC

The Adam optimiser's deep learning optimisation method was used to train the I-UNET technique, and eight epochs were used to track its validation and training loss metrics, as shown in Figure 6. From the first to the eighth epoch, the model continuously reduced loss levels, demonstrating an improved capacity to minimise discrepancies between expected outputs and actual targets. With an average loss level of about 1.53%, the predictions were accurate. This effective loss reduction was probably made possible by the Adam optimiser's adaptive learning rates and parameter updates, which allowed the model to converge to the best possible solution.

Fig. 6. Training and validation loss

4.3. Performance Metrics. The performance metrics include a variety of critical indications for assessing a model's success. Accuracy, AUC, and EER are among these measurements in equation (12-14).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}. \tag{12}$$

This metric provides an overall measure of the model's prediction accuracy, representing the ratio of successfully detected instances to total occurrences.

$$AUC = \sum \frac{(TPR[i] + TPR[i+1])}{2} * FPR[i+1] - FPR[i]). \tag{13}$$

AUC and the Receiver Operating Characteristics (ROC) curve are often associated. It shows the area under the curve, showing how the true positive and false positive rates are traded off at various classification levels. A higher AUC suggests better model performance overall. The EER metric is considered as follows:

$$EER = \frac{FP+FN}{TC}, \tag{14}$$

where TP is True Positive, TN is True Negative, FP denoted as False Positive, FN indicated as False Negative, TPR is True Positive rate, FPR represented as False positive rate.

TC displays the overall frames from the test dataset. As a result, a model that performs better has a greater AUC and a lower EER since it can discriminate more effectively. The following graph depicts the overall efficacy of video anomaly detection:

The existing methods such as the Two-stream fusion algorithm [31], AlexNet-based model [32], and Convolutional autoencoder [33] are compared with the proposed technique for analyzing the ROC curve. The Two-stream Fusion Algorithm (0.979) achieves good performance by combining temporal and spatial information. The ROC of AlexNet-based Model Convolutional Autoencoder are 0.970 and 0.9382. The proposed I-UNET is a significantly better U-Net model than the one that has been suggested, achieving flawless performance of ROC (1.00). Table 2 represents the performance of the ROC curve of the proposed model with existing works.

Table 2. Performance of ROC curve

Model	ROC curve
Two-stream fusion algorithm [31]	0.979
AlexNet-based model [32]	0.970
Convolutional autoencoder [33]	0.9382
I-UNET (Proposed)	1.00

Multiple performance measures are used to assess a model's effectiveness. Predicting accuracy requires understanding accuracy, which is defined as the proportion of successfully classified occurrences to all occurrences. AUC-ROC is a crucial metric for binary classification issues since it demonstrates the trade-off between true and false positive rates at various thresholds. The Equal Error Rate (EER) makes it easier to assess the model's performance objectively. This describes the point on the ROC curve when the rates of false rejection and mistaken acceptance are equal. These metrics evaluate a technique’s capacity to generate precise classifications over an extensive array of performance standards. Figure 7 represents the ROC curve with an actual positive rate and a false positive rate.

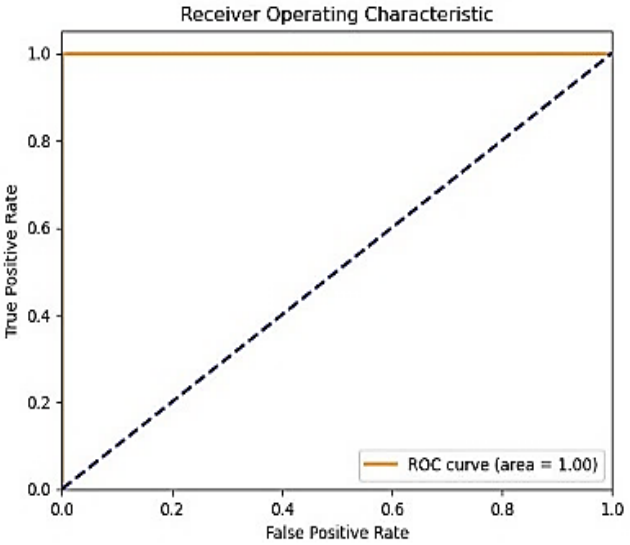


Fig. 7. ROC curve

4.4. Results Obtained. Figure 8 represents the final classification of video anomaly detection.

By adding anomaly scores, the I-UNET displays effective anomaly prediction. This model strategically integrates spatial and temporal information by employing a unique design that includes ConvLSTM layers. This new method improves the model's accuracy using encoder and decoder components that extract spatial and temporal data using ConvLSTM. Incorporating ConvLSTM enables the model to record detailed patterns across time, allowing it to recognise anomalies in video sequences more accurately. The cascade sliding window technique (CSWT) detects anomalies by calculating an anomaly score. This technique is critical in determining the existence or absence of abnormalities in each frame. It successfully analyses the successive frames, using a cascading sliding window technique to compute anomaly scores, thereby providing a dependable mechanism for identifying anomalies in video data.

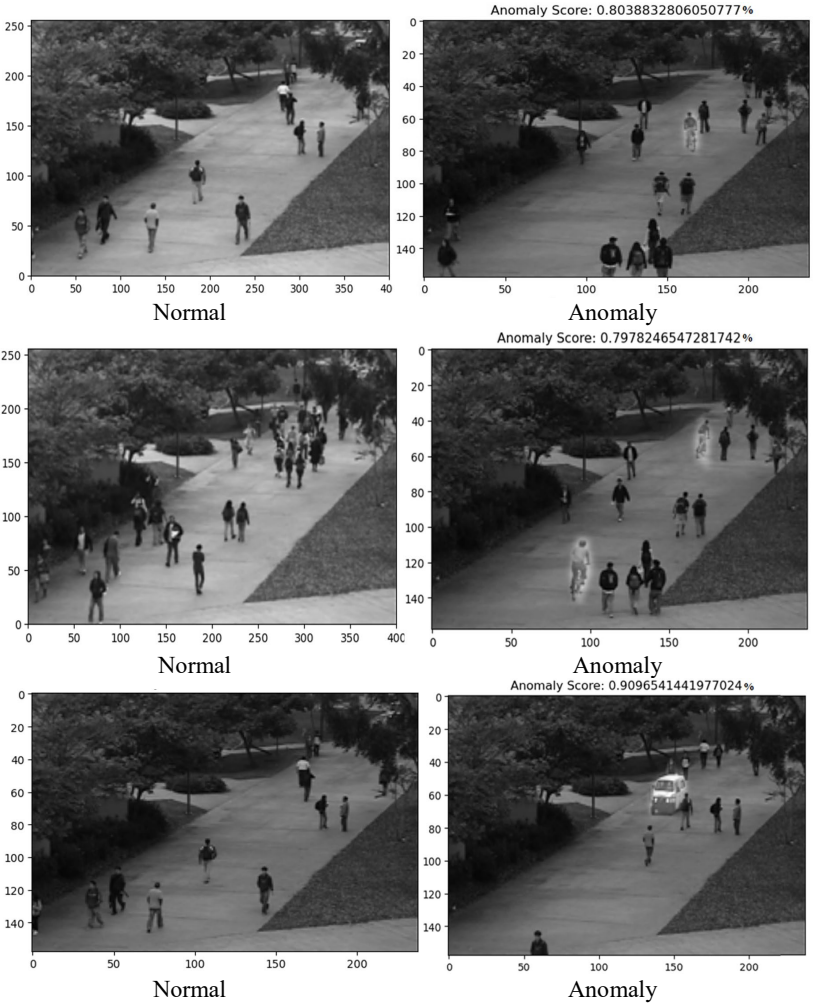


Fig. 8. Prediction result of image normal and anomaly image

4.5. Comparative Analysis. This section illustrates the recommended methodology's superior performance, with the innovative I-UNET serving as the fundamental framework. Compared to other models with fewer parameters, the unique I-UNET regularly outperforms, giving comparable or even greater performance. This highlights the efficiency of the proposed approach, establishing the I-UNET as a reliable and effective model for the given task. It combined the results of this model with those of

current research studies that specified strategies such as Approximated optical flow monitors algorithm (AOFM) [18], Mixture of Dynamic Texture (MDT) [18], Social Force SF) [18], Social Force and Mixture of Probabilistic Component Analyser (SF+MPPCA) [18], Hybrid Ensample Recurrent Reinforcement Model (HERR) [18], Mixture of probabilistic Principal component Analysis (MPPCA) [30], Convolution Auto-encoder (Conv-AE) [30], Convolution-Long short term memory-Auto-encoder (Conv-LSTM-AE) [30] unmasking [30].

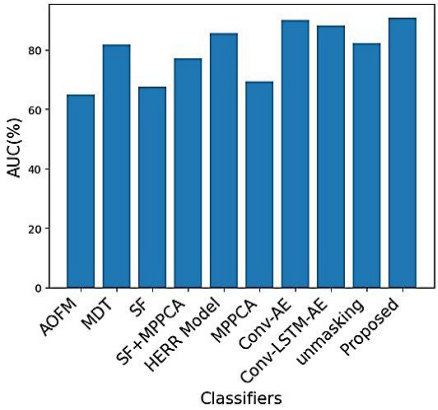


Fig. 9. Comparison of AUC

The research on the performance of different anomaly detection techniques emphasises the Area Under the Curve (AUC) measure. These AUC comparisons are shown in a figure, most likely Figure 9. The article emphasises how much better an innovative method known as I-UNET is at raising anomaly detection's AUC. The analysis shows that the accuracy of the I-UNET technique is higher than that of numerous conventional methods, such as AOFM, MDT, SF, SF+MPPCA, the HERR, MPPCA, Conv-AE, Conv-LSTM-AE, and unmasking model. According to these methodologies, the improvement percentages are as follows: 63%, 85%, 63%, 71%, 89.98%, 69.3%, 90.0%, 88.1%, and 82.2%. In terms of accuracy, conventional approaches continue to outperform the novel approach despite the remarkable performance increases attained by the I-UNET methodology. Although impressive, the AUC of 90.8% produced by the I-UNET methodology is not as high as that of existing methods.

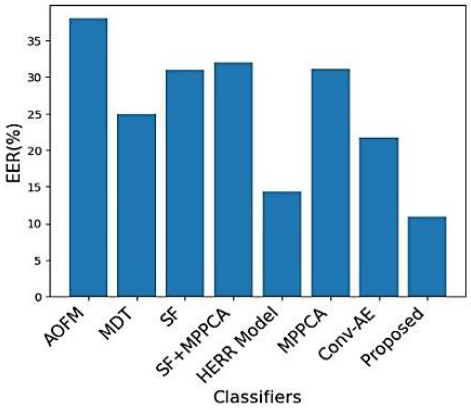


Fig. 10. Comparison of EER

The Equal Error Rate (EER) comparison of several anomaly detection techniques is shown in Figure 10. It demonstrates how well a brand-new strategy known as I-UNET works to lower anomaly detection's EER. The results show that the EER of the I-UNET approach is lower than that of numerous traditional approaches, such as AOFM, MDT, SF, SF+MPPCA, HERR, MPPCA, and Conv-AE model. The corresponding improvement percentages over these strategies are 42%, 25%, 31%, 32%, 14.33%, 31.1%, and 21.7%. Significantly, with an EER of 10.9%, the innovative I-UNET methodology outperforms conventional techniques in accuracy. This suggests a notable decrease in mistake rates in contrast to traditional methods. In addition, Table 3 offers a thorough summary of both AUC and EER values, enabling a more in-depth analysis of how well various approaches perform across these parameters.

Table 3. Comparisons of AUC and EER

Method	AUC(%)	EER(%)
AOFM [18]	63	42
MDT [18]	85	25
SF [18]	63	31
SF+MPPCA [18]	71	32
HERR model [18]	89.8	14.33
MPPCA [30]	69.3	31.1
Conv-AE [30]	90.0	21.7
Conv-LSTM-AE [30]	88.1	-
Unmasking [30]	82.2	-
I-UNET (Proposed)	90.8	10.9%

When assessing the effectiveness of a classification system, accuracy and Equal Error Rate (EER) are essential metrics, mainly when there are uneven class distributions or different sorts of errors. Although accuracy gauges forecasts' accuracy, it might give a partial picture in skewed conditions. By balancing the rates of false positives and false negatives, EER measures the model's performance. A lower EER indicates equal error costs and balanced class distribution produces better performance and positive correlations.

Compared to existing works such as AOFM, MDT, SF, SF+MPPCA, HERR, MPPCA, Conv-AE, Conv-LSTM-AE, and unmasking model, the proposed model attains high AUC and low EER. In this model, three datasets are used: UCF, UCSDped1, and UCSped2 datasets. Among these datasets, the UCSDped2 attains the highest accuracy and AUC but has a lower EER than other datasets. The performance analysis of metrics such as accuracy, AUC and EER for different datasets is represented in Table 4.

Table 4. Performance analysis of different datasets for the proposed model

Datasets	Accuracy (%)	AUC (%)	EER (%)
UCF	92	82.5	23.5
UCSDped1	96.5	87.8	18.6
UCSDped2	99	90.8	10.9

5. Conclusion. The researchers have focused on developing algorithms for reconstruction and prediction to tackle the complex problem of video anomaly identification in computer vision. Existing techniques encountered difficulties in unsupervised anomaly recognition due to resolution constraints and a lack of labelled anomalies, resulting in lesser accuracy. This paper presents a unique approach designated as Improved UNET (I-UNET) to reduce the risk of overfitting by addressing the need for sophisticated models capable of managing fine-grained data in video anomalies. A Weiner filter is used in the preprocessing stage to remove noise from video frames. The proposed architecture integrates spatial and temporal information in both the encoder and decoder sections by employing a ConvLSTM layer, ensuring anomaly detection precision. The Cascade Sliding Window Technique (CSWT) is used to calculate anomaly scores and assess the presence of anomaly frames to improve post-processing. The results show that the proposed network successfully segments anomalies, resulting in significantly better performance metrics such as Accuracy of 99%, AUC of 90.8%, and EER of 10.9%. This demonstrates the efficacy of the suggested methodology in detecting high-precision video anomalies, a significant advancement in the field. Fine-

tuning and adaptation are crucial in tailoring pre-trained models to specific anomaly detection tasks. This process involves optimising hyperparameters, implementing regularisation techniques, and devising effective adaptation strategies. The future potential of utilising transfer learning techniques with pre-trained models for anomaly detection on comparable tasks or datasets warrants examination. This approach holds promise for identifying anomalies in scenarios where labelled data is limited or unavailable.

References

1. Ramachandra B., Jones M.J., Vatsavai R.R. A survey of single-scen4e video anomaly detection. IEEE transactions on pattern analysis and machine intelligence. 2020. vol. 44(5). pp. 2293–2312.
2. Nayak R., Pati U.C., Das S.K. A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing. 2021. vol. 106(6). DOI: 10.1016/j.imavis.2020.104078.
3. Raja R., Sharma P.C., Mahmood M.R., Saini D.K. Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimedia Tools and Applications. 2023. vol. 82(8). pp. 12635–12651.
4. Erhan L., Ndubuaku M., Di Mauro M., Song W., Chen M., Fortino G., Bagdasar O., Liotta A. Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion. 2021. vol. 67. pp. 64–79.
5. Pang G., Shen C., Cao L., Hengel A.V.D. Deep learning for anomaly detection: A review. ACM computing surveys (CSUR). 2021. vol. 54(2). pp. 1–38.
6. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2024. vol. 28(1). pp. 135–151.
7. Ackerson J.M., Dave R., Seliya N. Applications of recurrent neural network for biometric authentication & anomaly detection. Information. 2021. vol. 12(7). DOI: 10.3390/info12070272.
8. Şengönül E., Samet R., Abu Al-Haija Q., Alqahtani A., Alturki B., Alsulami A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. Applied Sciences. 2023. vol. 13(8). DOI: 10.3390/app13084956.
9. da Costa K.A., Papa J.P., Passos L.A., Colombo D., Del Ser J., Muhammad K., de Albuquerque V.H.C. A critical literature survey and prospects on tampering and anomaly detection in image data. Applied Soft Computing. 2020. vol. 97. DOI: 10.1016/j.asoc.2020.106727.
10. Jebur S.A., Hussein K.A., Hoomod H.K., Alzubaidi L., Santamaría J. Review on deep learning approaches for anomaly event detection in video surveillance. Electronics. 2022. vol. 12(1). DOI: 10.3390/electronics12010029.
11. Habeeb R.A.A., Nasaruddin F., Gani A., Hashem I.A.T., Ahmed E., Imran M. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management. 2019. vol. 45. pp. 289–307.
12. Arshad K., Ali R.F., Muneer A., Aziz I.A., Naseer S., Khan N.S., Taib S.M. Deep Reinforcement Learning for Anomaly Detection: A Systematic Review. IEEE Access. 2022. vol. 10. pp. 124017–124035.
13. Berroukham A., Housni K., Lahraichi M., Boulfrifi I. Deep learning-based methods for anomaly detection in video surveillance: a review. Bulletin of Electrical Engineering and Informatics. 2023. vol. 12(1). pp. 314–327.

14. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4(2). DOI: 10.3390/jimaging4020036.
15. Musa A.A., Hussaini A., Liao W., Liang F., Yu W. Deep Neural Networks for Spatial-Temporal Cyber-Physical Systems: A Survey. Future Internet. 2023. vol. 15(6). DOI: 10.3390/fi15060199.
16. Albuquerque Filho J.E., Brandão L.C., Fernandes B.J., Maciel A.M. A review of neural networks for anomaly detection. IEEE Access. 2022. vol. 10(5). pp. 112342–112367.
17. Borowiec M.L., Dikow R.B., Frandsen P.B., McKeeken A., Valentini G., White A.E. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution. 2022. vol. 13(8). pp. 1640–1660.
18. Amudha L., Pushpa Lakshmi R. Performance Analysis of Hybrid RR Algorithm for Anomaly Detection in Streaming Data. Computer Systems Science & Engineering. 2023. vol. 45(3). pp. 2299–2312.
19. Chang Y., Tu Z., Xie W., Luo B., Zhang S., Sui H., Yuan J. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition. 2022. vol. 122. DOI: 10.1016/j.patcog.2021.108213.
20. Rezaei F., Yazdi M. A new semantic and statistical distance-based anomaly detection in crowd video surveillance. Wireless Communications and Mobile Computing. 2021. vol. 2021. DOI: 10.1155/2021/5513582.
21. Deepak K., Chandrakala S., Mohan C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 2021. vol. 15(1). pp. 215–222.
22. Ul Amin S., Ullah M., Sajjad M., Cheikh F.A., Hijji M., Hijji A., Muhammad K. EADN: An efficient deep learning model for anomaly detection in videos. Mathematics. 2022. vol. 10(9). DOI: 10.3390/math10091555.
23. Taghinezhad N., Yazdi M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction. IEEE Access. 2023. vol. 11. pp. 9295–9310.
24. Liu T., Zhang C., Niu X., Wang L. Spatio-temporal prediction and reconstruction network for video anomaly detection. Plos one. 2022. vol. 17(5). DOI: 10.1371/journal.pone.0265564.
25. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
26. Chriki A., Touati H., Snoussi H., Kamoun F. Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications. 2021. vol. 80. pp. 2599–2620.
27. Deepak K., Srivathsan G., Roshan S., Chandrakala S. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. Circuits, Systems, and Signal Processing. 2021. vol. 40(3). pp. 1333–1349.
28. dos Santos J.C.M., Carrijo G.A., de Fátima dos Santos Cardoso C., Ferreira J.C., Sousa P.M., Patrocínio A.C. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. Research on Biomedical Engineering. 2020. vol. 36. pp. 107–119.
29. Sharma N., Gupta S., Koundal D., Alyami S., Alshahrani H., Asiri Y., Shaikh A. U-Net model with transfer learning model as a backbone for segmentation of gastrointestinal tract. Bioengineering. 2023. vol. 10(1). DOI: 10.3390/bioengineering10010119.
30. Cai Y., Liu J., Guo Y., Hu S., Lang S. Video anomaly detection with multi-scale feature and temporal information fusion. Neurocomputing. 2021. vol. 423. pp. 264–273.

31. Yang Y., Fu Z., Naqvi S.M. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing. 2023. vol. 553. DOI: 10.1016/j.neucom.2023.126561.
32. Khan A.A., Nauman M.A., Shoaib M., Jahangir R., Alroobaea R., Alsafyani M., Binmahfoudh A., Wechtaisong C. Crowd anomaly detection in video frames using fine-tuned AlexNet Model. Electronics. 2022. vol. 11(19). DOI: 10.3390/electronics11193105.
33. Ali M.M. Real-time video anomaly detection for smart surveillance. IET Image Processing. 2023. vol. 17(5). pp. 1375–1388.

R. Krishnan Sreedevi — Research scholar, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women. Research interests: deep learning, computer vision, machine learning, network security and anomaly detection. The number of publications — 6. 19pheop005@avinuty.ac.in; Tamil Nadu, 641043, Coimbatore, India; office phone: +91(960)535-9348.

Amudha P. — Professor, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women. Research interests: data mining, machine learning, information security. The number of publications — 56. amudha_cse@avinuty.ac.in; Tamil Nadu, 641043, Coimbatore, India; office phone: +91(902)563-6594.

УДК 004.8 **DOI** 10.15622/ia.23.6.12

Ш. Р. Кришнан, П. Амудха

УЛУЧШЕНИЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ НА ВИДЕО С ПОМОЩЬЮ УСОВЕРШЕНСТВОВАННОЙ ТЕХНОЛОГИИ UNET И ТЕХНИКИ КАСКАДНОГО СКОЛЬЗЯЩЕГО ОКНА

Р. Кришнан Ш., Амудха П. **Улучшение обнаружения аномалий на видео с помощью усовершенствованной технологии UNET и техники каскадного скользящего окна.**

Аннотация. Обнаружение аномалий на видео с помощью компьютерного зрения все еще нуждается в совершенствовании, особенно при распознавании изображений с необычными движениями или объектами. Современные подходы в основном сосредоточены на методах реконструкции и прогнозирования, а обнаружение аномалий на видео без наблюдения сталкивается с трудностями из-за отсутствия достаточного количества помеченных аномалий, что снижает точность. В этой статье представлена новая структура под названием усовершенствованная UNET (I-UNET), разработанная для противодействия переобучению путем удовлетворения потребности в сложных моделях, которые могут извлекать малозаметную информацию из аномалий на видео. Видеошум можно устранить путем предварительной обработки кадров фильтром Винера. Более того, система использует сверточные слои долго-кратковременной памяти (ConvLSTM) для плавной интеграции временных и пространственных данных в свои части энкодера и декодера, улучшая точность идентификации аномалий. Последующая обработка осуществляется с использованием техники каскадного скользящего окна (CSWT) для идентификации аномальных кадров и генерации оценок аномалии. По сравнению с базовыми подходами, экспериментальные результаты на наборах данных UCF, UCSDped1 и UCSDped2 демонстрируют заметные улучшения производительности, с точностью 99%, площадью под кривой (AUC) 90,8% и равным уровнем ошибок (EER) 10,9%. Это исследование предоставляет надежную и точную структуру для обнаружения аномалий на видео с наивысшим уровнем точности.

Ключевые слова: обнаружение аномалий, I-UNET, фильтр Винера, ConvLSTM, каскадное скользящее окно, оценка аномалий.

Литература

1. Ramachandra B., Jones M.J., Vatsavai R.R. A survey of single-scen4e video anomaly detection. IEEE transactions on pattern analysis and machine intelligence. 2020. vol. 44(5). pp. 2293–2312.
2. Nayak R., Pati U.C., Das S.K. A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing. 2021. vol. 106(6). DOI: 10.1016/j.imavis.2020.104078.
3. Raja R., Sharma P.C., Mahmood M.R., Saini D.K. Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimedia Tools and Applications. 2023. vol. 82(8). pp. 12635–12651.
4. Erhan L., Ndubuaku M., Di Mauro M., Song W., Chen M., Fortino G., Bagdasar O., Liotta A. Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion. 2021. vol. 67. pp. 64–79.
5. Pang G., Shen C., Cao L., Hengel A.V.D. Deep learning for anomaly detection: A review. ACM computing surveys (CSUR). 2021. vol. 54(2). pp. 1–38.
6. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2024. vol. 28(1). pp. 135–151.

7. Ackerson J.M., Dave R., Seliya N. Applications of recurrent neural network for biometric authentication & anomaly detection. Information. 2021. vol. 12(7). DOI: 10.3390/info12070272.
8. Şengönül E., Samet R., Abu Al-Haija Q., Alqahtani A., Alturki B., Alsulami A.A. An Analysis of Artificial Intelligence Techniques in Surveillance Video Anomaly Detection: A Comprehensive Survey. Applied Sciences. 2023. vol. 13(8). DOI: 10.3390/app13084956.
9. da Costa K.A., Papa J.P., Passos L.A., Colombo D., Del Ser J., Muhammad K., de Albuquerque V.H.C. A critical literature survey and prospects on tampering and anomaly detection in image data. Applied Soft Computing. 2020. vol. 97. DOI: 10.1016/j.asoc.2020.106727.
10. Jebur S.A., Hussein K.A., Hoomod H.K., Alzubaidi L., Santamaría J. Review on deep learning approaches for anomaly event detection in video surveillance. Electronics. 2022. vol. 12(1). DOI: 10.3390/electronics12010029.
11. Habeeb R.A.A., Nasaruddin F., Gani A., Hashem I.A.T., Ahmed E., Imran M. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management. 2019. vol. 45. pp. 289–307.
12. Arshad K., Ali R.F., Muneer A., Aziz I.A., Naseer S., Khan N.S., Taib S.M. Deep Reinforcement Learning for Anomaly Detection: A Systematic Review. IEEE Access. 2022. vol. 10. pp. 124017–124035.
13. Berroukham A., Housni K., Lahraichi M., Boulfrifi I. Deep learning-based methods for anomaly detection in video surveillance: a review. Bulletin of Electrical Engineering and Informatics. 2023. vol. 12(1). pp. 314–327.
14. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4(2). DOI: 10.3390/jimaging4020036.
15. Musa A.A., Hussaini A., Liao W., Liang F., Yu W. Deep Neural Networks for Spatial-Temporal Cyber-Physical Systems: A Survey. Future Internet. 2023. vol. 15(6). DOI: 10.3390/fi15060199.
16. Albuquerque Filho J.E., Brandão L.C., Fernandes B.J., Maciel A.M. A review of neural networks for anomaly detection. IEEE Access. 2022. vol. 10(5). pp. 112342–112367.
17. Borowiec M.L., Dikow R.B., Frandsen P.B., McKeeken A., Valentini G., White A.E. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution. 2022. vol. 13(8). pp. 1640–1660.
18. Amudha L., Pushpa Lakshmi R. Performance Analysis of Hybrid RR Algorithm for Anomaly Detection in Streaming Data. Computer Systems Science & Engineering. 2023. vol. 45(3). pp. 2299–2312.
19. Chang Y., Tu Z., Xie W., Luo B., Zhang S., Sui H., Yuan J. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition. 2022. vol. 122. DOI: 10.1016/j.patcog.2021.108213.
20. Rezaei F., Yazdi M. A new semantic and statistical distance-based anomaly detection in crowd video surveillance. Wireless Communications and Mobile Computing. 2021. vol. 2021. DOI: 10.1155/2021/5513582.
21. Deepak K., Chandrakala S., Mohan C.K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 2021. vol. 15(1). pp. 215–222.
22. Ul Amin S., Ullah M., Sajjad M., Cheikh F.A., Hijji M., Hijji A., Muhammad K. EADN: An efficient deep learning model for anomaly detection in videos. Mathematics. 2022. vol. 10(9). DOI: 10.3390/math10091555.

23. Taghinezhad N., Yazdi M. A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction. IEEE Access. 2023. vol. 11. pp. 9295–9310.
24. Liu T., Zhang C., Niu X., Wang L. Spatio-temporal prediction and reconstruction network for video anomaly detection. Plos one. 2022. vol. 17(5). DOI: 10.1371/journal.pone.0265564.
25. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
26. Chriki A., Touati H., Snoussi H., Kamoun F. Deep learning and handcrafted features for one-class anomaly detection in UAV video. Multimedia Tools and Applications. 2021. vol. 80. pp. 2599–2620.
27. Deepak K., Srivathsan G., Roshan S., Chandrakala S. Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders. Circuits, Systems, and Signal Processing. 2021. vol. 40(3). pp. 1333–1349.
28. dos Santos J.C.M., Carrijo G.A., de Fátima dos Santos Cardoso C., Ferreira J.C., Sousa P.M., Patrocínio A.C. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. Research on Biomedical Engineering. 2020. vol. 36. pp. 107–119.
29. Sharma N., Gupta S., Koundal D., Alyami S., Alshahrani H., Asiri Y., Shaikh A. U-Net model with transfer learning model as a backbone for segmentation of gastrointestinal tract. Bioengineering. 2023. vol. 10(1). DOI: 10.3390/bioengineering10010119.
30. Cai Y., Liu J., Guo Y., Hu S., Lang S. Video anomaly detection with multi-scale feature and temporal information fusion. Neurocomputing. 2021. vol. 423. pp. 264–273.
31. Yang Y., Fu Z., Naqvi S.M. Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing. 2023. vol. 553. DOI: 10.1016/j.neucom.2023.126561.
32. Khan A.A., Nauman M.A., Shoaib M., Jahangir R., Alroobaea R., Alsafyani M., Binmahfoudh A., Wechtaisong C. Crowd anomaly detection in video frames using fine-tuned AlexNet Model. Electronics. 2022. vol. 11(19). DOI: 10.3390/electronics11193105.
33. Ali M.M. Real-time video anomaly detection for smart surveillance. IET Image Processing. 2023. vol. 17(5). pp. 1375–1388.

Р. Кришнан Шридеви — научный сотрудник, кафедра компьютерных наук и инженерии, Институт домоводства и высшего образования для женщин Авинашилингам. Область научных интересов: глубокое обучение, компьютерное зрение, машинное обучение, сетевая безопасность и обнаружение аномалий. Число научных публикаций — 6. 19pheop005@avinuty.ac.in; Тамил Наду, 641043, Коимбатур, Индия; р.т.: +91(960)535-9348.

Амудха П. — профессор, кафедра компьютерных наук и инженерии, Институт домоводства и высшего образования для женщин Авинашилингам. Область научных интересов: интеллектуальный анализ данных, машинное обучение, информационная безопасность. Число научных публикаций — 56. amudha_cse@avinuty.ac.in; Тамил Наду, 641043, Коимбатур, Индия; р.т.: +91(902)563-6594.

ПАМЯТИ ЮСУПОВА РАФАЭЛЯ МИДХАТОВИЧА

7 ноября 2024 года, на 91-м году ушел из жизни Рафаэль Мидхатович Юсупов – выдающийся ученый в области информатики, информационных технологий и теории управления, основатель и руководитель научных школ по теоретическим основам информатизации общества и по теории чувствительности сложных информационно-управляющих систем, доктор технических наук, профессор, член-корреспондент РАН, заслуженный деятель науки и техники Российской Федерации, руководитель научного направления СПИИРАН Санкт-Петербургского Федерального исследовательского центра Российской академии наук, директор СПИИРАН (1991-2018 гг.).

Юсупов Рафаэль Мидхатович родился 17 июля 1934 г. в г. Казани. После окончания в 1952 г. с золотой медалью Казанской спецшколы ВВС Р.М. Юсупов был направлен в Ленинградскую военно-воздушную инженерную академию (ныне Военно-космическая академия имени А.Ф. Можайского), которую окончил с отличием в 1958 г. по специальности «инженер-электрик». В 1964 г. он окончил Ленинградский государственный университет по специальности «математика».

В 1958-1985 гг. Р.М. Юсупов проходил службу в Военной академии имени А.Ф. Можайского (ВА имени А.Ф. Можайского), где занимал должности инженера (1958-1959 гг.), старшего инженера (1959 г.), научного сотрудника (1959-1960 гг.), адъюнкта (1960-1962 гг.), старшего научного сотрудника (1962-1967 гг.), начальника научно-исследовательской лаборатории систем управления летательных аппаратов (1967-1970 гг.), заместителя начальника кафедры систем управления ракет и космических аппаратов (1970-

1971 гг.), начальника кафедры боевой эффективности применения ракетно-космической техники (1971-1976 гг.), начальника созданного с его участием, уникального в системе высшего военного образования, факультета сбора и обработки информации (1976-1985 гг.). В 1985 г. Р.М. Юсупов назначается на должность начальника Направления моделирования стратегических операций Центра оперативно-стратегических исследований Генерального Штаба Вооруженных Сил СССР. В 1986 г. с целью улучшения оперативно-тактической подготовки выпускников ВА имени А.Ф. Можайского для службы в космических войсках, учитывая опыт служебной деятельности Р.М. Юсупова в ГШ ВС, он был назначен начальником кафедры оперативно-тактической подготовки ВА имени А.Ф. Можайского.

В 1989 г. Р.М. Юсупов назначен с оставлением в кадрах Вооруженных Сил заместителем директора по научной работе Ленинградского института информатики и автоматизации АН ССС (с 1991 г. – Санкт-Петербургский институт информатики и автоматизации), а в 1991 г. избран директором этого института и проработал в этой должности до 2018 г. С 2018 г. Р.М. Юсупов являлся руководителем научного направления СПИИРАН.

Научно-педагогическая деятельность Р.М. Юсупова началась (с 1959 г.) с должности научного сотрудника ВКА, а впоследствии включала должности начальника военных кафедр и факультета, и заведующего кафедрами, и декана факультетов гражданских ВУЗов до директора Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), одного из ведущих институтов РАН в области информатики и автоматизации.

Р.М. Юсупов – организатор, руководитель и участник крупнейших фундаментальных и прикладных исследований в области кибернетики и информатики, результаты которых явились важным вкладом в отечественную и мировую науку, образование и экономику страны, ее безопасность и развитие международных научных связей.

В 1958-1976 гг. Р.М. Юсупов проводил исследования в области теории управления полетом летательных аппаратов, теории самонастраивающихся (адаптивных) систем, теории идентификации и теории чувствительности динамических систем. Он стал одним из основоположников отечественной научной школы по теории чувствительности. Благодаря работам Р.М. Юсупова, его коллег и учеников теория чувствительности вошла в число основных разделов общего курса теории автоматического управления. Анализ чувствительности стал обязательным этапом создания систем управления и их элементов. Научные результаты Р.М. Юсупова в

указанных областях используются при проектировании, настройке и испытаниях высокоточных приборов и систем.

В 1981 г. в издательстве «Наука» по теме диссертации совместно с Е.Н. Розенвассером опубликована монография «Чувствительность систем управления», ее актуальность подтверждается публикацией в США английского перевода этой монографии.

С 1976 г. научные интересы Р.М. Юсупова связаны с проблемами сбора и обработки информации, геофизической кибернетики, математического моделирования, информатизации, телемедицины, конфликтологии, информационной безопасности.

Термин геофизическая кибернетика был предложен Р.М. Юсуповым в 1977 г. применительно к новому междисциплинарному научному направлению – теории и практике управления природными (геофизическими) объектами и процессами.

Под его руководством и с его участием в ВКА им. А.Ф. Можайского и в СПИИРАН выполнено, в том числе по постановлениям Правительства и Федеральным целевым научно-техническим программам, более 90 научно-исследовательских работ и проектов по вопросам повышения точности, эффективности, испытаний военно-технических систем (ВТС), планирования их развития, создания математического обеспечения систем обработки информации в ВТС, разработки информационно-расчетных систем оценки состояния природной среды и ее влияния на применение ВТС, создания новых информационных технологий и их использования в различных сферах развития общества.

В области теории моделирования Р.М. Юсупов развил новое научное направление – методы оценивания качества моделей (адекватность, чувствительность, сложность и т.д.), названное им квалиметрией моделей (моделеметрией). Им сформулированы концептуальные основы квалиметрии моделей, разработаны методы и алгоритмы оценивания адекватности и чувствительности моделей.

Избрание Р.М. Юсупова в 1991 г. директором СПИИРАН совпало с массовым переходом во всем мире и в нашей стране на новое поколение вычислительной техники – персональные компьютеры. Персональные компьютеры приблизили мощные вычислительные ресурсы к непосредственному потребителю и проникли во все сферы человеческой деятельности. При этом стало ясно, что простая компьютеризация не может обеспечить обществу прорыва на новый технологический уровень. Решением этой проблемы стал переход от простой компьютеризации к более широкой и глубокой информатизации общества.

Р.М. Юсупов явился одним из инициаторов развития в России взамен исчерпавшего себя направления – компьютеризация нового весьма актуального научно-практического направления – информатизация общества. Разработанная им универсальная структура концепции информатизации, по существу, стала в России основой всех работ в области информатизации, чему, в частности, способствовали опубликованные им совместно с В.П. Заболотским монографии «Научно-методологические основы информатизации», СПб, Наука, 2001 и «Концептуальные и научно-методологические основы информатизации», СПб, Наука, 2009.

Впервые эти концептуальные и научно-методологические основы информатизации были реализованы на практике в 1991-1993 гг., когда под руководством Р.М. Юсупова и с его непосредственным участием были созданы концепция информатизации Ленинградского экономического региона (1991 г.) и типовая концепция информатизации городского района (1992 г). Р.М. Юсупов является соавтором Концепции информатизации Санкт-Петербурга. С его участием разработана также стратегия перехода Санкт-Петербурга в информационное общество, он является руководителем и разработчиком ряда программ и приоритетных проектов для города в области телекоммуникаций, информационной безопасности, телемедицины, в том числе Целевой программы «Электронный Санкт-Петербург», Концепции информационной безопасности исполнительных органов государственной власти Санкт-Петербурга и др.

Р.М. Юсупов являлся бессменным председателем оргкомитета конференций, проводимых в городе на регулярной основе при поддержке Правительства Санкт-Петербурга: «Региональная информатика» (1991-2024 гг.) и «Информационная безопасность регионов России» (1999-2023 гг.), которые способствуют формированию единого защищенного информационного пространства и информационно-образовательной среды города. Им был организован постоянно действующий семинар по разработке компьютерных моделей развития города и городской семинар при Научном совете по информатизации Санкт-Петербурга «Информатика и автоматизация». Благодаря его усилиям, СПИИРАН стал научно-методическим центром информатизации Санкт-Петербурга.

Значительные результаты, полученные Р.М. Юсуповым при разработке концептуальных и научно-методологических основ информатизации и информационного общества, оказали существенное влияние на эти процессы в Санкт-Петербурге и стране. В развитие этого направления Р.М. Юсуповым разработаны структурная

и экономико-математическая модели информационного общества в предположении, что в последнем функционируют секторы материального и информационного (цифрового) производства. С использованием этих моделей им получена новая параметрическая информационная модель развития науки, позволяющая исследовать влияние ряда параметров (объем финансирования, численность научных работников, старение знаний, «утечка умов» и технологий и т.д.) на эволюцию науки.

Юсуповым Р.М. созданы две прикладные теории управления, имеющие важное государственное и оборонное значение в решении проблем импортозамещения и импортоопережения в различных отраслях современной экономики РФ. Первая из указанных теорий получила название теории проактивного (упреждающего) управления сложными объектами, а вторая, дополняющая ее теория, была названа теорией многокритериального оценивания и выбора наиболее предпочтительных моделей и полимодельных комплексов, описывающих функционирование сложных объектов. Теории вносят существенный вклад в развитие современной информатики и ее составной части – искусственного интеллекта.

В результате реализации этих теорий решена крупная научно-техническая проблема обеспечения технологической независимости российских разработчиков от зарубежных производителей в области проектирования, создания и эксплуатации модельно-алгоритмического, технического, информационного и программного обеспечения проактивного управления жизненным циклом сложных военно-технических объектов и под руководством и непосредственном участии Юсупова Р.М. разработана отечественная интеллектуальная информационно-аналитическая платформа проактивного управления жизненным циклом сложных военно-технических объектов на основе применения киберфизических систем и интеллектуальных интерфейсов.

Свою высокоэффективную научную деятельность Р.М. Юсупов всегда успешно сочетал с работой по подготовке высококвалифицированных инженерных и научных кадров. В течение 20 лет в ВКА имени А.Ф. Можайского на должностях начальника кафедры и начальника факультета он готовил высококвалифицированные инженерные кадры для Вооруженных Сил, многие из которых в настоящее время с успехом трудятся в различных структурах, в том числе и в РАН.

С 1991 г. он являлся заведующим базовой кафедрой «Автоматизация научных исследований» Санкт-Петербургского государственного электротехнического университета «ЛЭТИ»,

а с 2003 по 2016 гг. заведовал базовой кафедрой «Прикладная информатика» Санкт-Петербургского государственного университета аэрокосмического приборостроения.

В 1996-1999 гг. Р.М. Юсупов был профессором Санкт-Петербургского политехнического университета Петра Великого (СПбПУ), где при его непосредственном участии в 1998 г. был создан факультет «Безопасность». Деканом этого факультета Р.М. Юсупов был с 1998 по 2001 гг. В 2009 г. он был избран заведующим базовой кафедрой «Распределённые интеллектуальные системы автоматизации» СПбПУ.

Р.М. Юсупов – почетный доктор Петрозаводского государственного университета, Санкт-петербургского университета управления и экономики, почетный профессор Военно-космической академии имени А.Ф. Можайского, почетный академик Академии наук Республики Татарстан.

По инициативе Р.М. Юсупова в здании института информатики созданы два музея. В 1995 г. открылся Музей истории всемирно известной школы К. Мая, которая с 1910 по 1976 гг. размещалась в здании института. С 2008 г. функционирует Музей истории СПИИРАН.

Р.М. Юсупов – выдающийся ученый в области информатики и теории управления, основатель и руководитель научных школ по теории чувствительности информационно-управляющих систем, квалиметрии моделей, геофизической кибернетике и научной школы «информатизация и формирование информационного общества». Среди учеников Р.М. Юсупова 12 докторов наук и 46 кандидатов наук. Р.М. Юсупов является автором более 500 научных трудов, 40 монографий, учебников и учебных пособий, 19 изобретений.

Р.М. Юсупов вел большую научно-организаторскую и общественную работу как член Президиума Санкт-Петербургского научного центра РАН (1992-2018 гг.), председатель Объединенного научного совета этого центра по информатике, телекоммуникациям и управлению (с 1992 г.), член Бюро Отделения нанотехнологий и информационных технологий РАН (до 2016 г.), член Научного совета при Совете Безопасности РФ (1999-2014 гг.), заместитель председателя Научного совета по информатизации Санкт-Петербурга (с 1994 г.), член Совета РАН «Научные телекоммуникации и информационная инфраструктура» (1998-2003 гг.), член Совета РАН «Высокопроизводительные вычислительные системы, научные телекоммуникации и информационная инфраструктура» (с 2003 г.), председатель Научно-технического совета по региональным

проблемам национальной безопасности (1996-1998 г.), заместитель председателя Научного совета при по информатизации Санкт-Петербурга при Правительстве города (1993-2024 гг.), член научных советов по государственным НТП «Перспективные информационные технологии» (1994-1998 гг.) и «Информатизация России» (с 1994 г.), член Северо-Западной секции содействия развитию экономической науки РАН (с 1998 г.), член комиссии при Губернаторе Санкт-Петербурга по реформированию научно-технической сферы (1998-1999 гг.), член правления Санкт-Петербургского отделения Ломоносовского фонда (с 2004 г.), член Общественного Совета Санкт-Петербурга (2002-2010 гг.), сопредседатель Координационного совета Партнерства для развития информационного общества на Северо-Западе России (ПРИОР Северо-Запад) (с 2002 г.), президент Национального общества имитационного моделирования (с 2011 г.), заместитель председателя Общественного совета Комитета по информатизации и связи Правительства Санкт-Петербурга (с 2014 г.), вице-президент Санкт-Петербургского Общества информатики, вычислительной техники, систем связи и управления (с 1991 г.), член Объединённого совета по прикладным наукам и технологическому развитию промышленности Санкт-Петербургского отделения Российской академии наук (с 2024 г.).

Являясь членом Научного совета при Совете Безопасности Российской Федерации, Р.М. Юсупов принимал участие в разработке более десяти проектов государственных документов по обеспечению информационной безопасности, в которые вошли его предложения. В одном из последних документов совета «Основные направления государственной политики в области формирования культуры информационной безопасности личности до 2020 года» учтены два его предложения. Являясь заместителем председателя Научного совета по информатизации Санкт-Петербурга, Р.М. Юсупов, опираясь на разработанные им концептуальные и методологические основы информатизации и развития информационного общества, внес важный вклад в разработку документов, оказавших существенное положительное влияние на эти процессы в Санкт-Петербурге и в стране.

Р.М. Юсупов входит в состав редакционных советов журналов: «Информатика и ее применения», «Научно-технические ведомости СПбПУ» (председатель редакционного совета серии «Информатика. Телекоммуникации. Управление»), «Информация и космос», «Известия Петербургского университета путей сообщения», «Экономика и управление», «Прикаспийский журнал. Управление и высокие технологии», «Проблемы управления и информатики» (Украина),

«Мехатроника, автоматизация и управление», «Информатизация и связь», «Телекоммуникации», «Journal of Intelligent Control» (США) др., много лет являлся бессменным председателем докторского диссертационного совета при СПИИРАН, членом докторских диссертационных советов при ВКА имени А.Ф. Можайского и СПбПУ.

Благодаря усилиям Р.М. Юсупова, как главного редактора, журнал «Информатика и автоматизация» («Труды СПИИРАН»), издаваемый институтом с 2001 г., включен в список ВАК с 2011 г., с 2014 г. вошел в международную базу данных Scopus, в 2018 г. включен в топ 100 российских журналов.

Международным признанием научных достижений Р.М. Юсупова являются присуждение ему ряда международных премий, его многочисленные научные публикации (в том числе монографии) в зарубежных издательствах, приглашение в состав программных и организационных комитетов международных конференций (в том числе и в качестве докладчика), выступления с приглашенными докладами на международных конференциях, чтение лекций в зарубежных университетах. Р.М. Юсупов избран членом ряда международных академий, входит в состав редколлегий ряда зарубежных журналов. Он является руководителем многих заказных зарубежных проектов и грантов. За работы в области обеспечения международной информационной безопасности он награжден орденом «Содружество» Межпарламентской ассамблеей государств-участников Содружества независимых государств (2013 г.), Почётным знаком МПА СНГ «За заслуги в области печати и информации» (2018 г.). В 2005 г. ему присуждена Международная премия им. Н. Рериха за достижения в области педагогики и просветительства.

За заслуги в период военной службы Р.М. Юсупов награжден орденом «Красной звезды» (1978 г.) и тринадцатью медалями, ему присвоено воинское звание генерал-майор (1980 г.). За заслуги в развитии военной науки, вооружения и техники он избран действительным членом Академии военных наук России (1996 г.), награждён грамотой Военно-научного комитета Вооруженных Сил РФ «За большой вклад в развитие отечественной науки, вооружения, и военной техники» (2012 г.).

За научные достижения и педагогические заслуги Р.М. Юсупов награжден орденом «За заслуги перед Отечеством» IV степени (2005 г.), орденом «Почета» (1999 г.), медалью Совета Безопасности РФ «За заслуги в обеспечении национальной безопасности» (2009 г.), ему присвоено почетное звание «Заслуженный деятель науки

и техники РФ» (1984 г.), «Почетный радист СССР» (1974 г.), присуждена ученая степень доктора технических наук (1968 г.), присвоено ученое звание профессора (1974 г.), он награжден Министерством высшего и среднего специального образования СССР Первой премией и медалью за лучшую научную работу (1983 г.), избран членом-корреспондентом Российской академии наук (2006 г.), награжден Почетной грамотой Президента РФ (2015 г.), удостоен звания «Почетный работник науки и высоких технологий Российской Федерации» за значительные заслуги в сфере науки и многолетний добросовестный труд (Приказ Минобрнауки № 38/к-н от 26 июня 2019 г.).

Он удостоен премий Правительства РФ и Правительства Санкт-Петербурга в области образования (2009 г.), премии Правительства Санкт-Петербурга им. А.С. Попова в области электро- и радиотехники и информационных технологий, награжден Почетным знаком «За заслуги перед Санкт-Петербургом» (2018 г.), Знаком отличия «За заслуги перед Санкт-Петербургом» (2009 г.).

Р.М. Юсупов являлся руководителем работы, удостоенной премии Правительства Российской Федерации в области науки и техники (2022 г.) – за разработку и внедрение комплекса отечественных интеллектуальных наземных транспортно-технологических средств обслуживания судов гражданской авиации в едином цифровом пространстве аэропорта.

Решением Президиума Санкт-Петербургского отделения РАН в октябре 2024 г. Рафаэлю Мидхатовичу присуждена Премия имени С.Н. Ковалёва за выдающиеся научные и научно-технические достижения в области технических наук.

За существенный вклад в научно-технологическое развитие Российской Федерации и содействие Российской академии наук в решении возложенных на нее задач Юсупов Р.М. награждён медалью 300 лет Российской академии наук (2024 г.), благодарственным письмом Президента Российской Федерации за вклад в развитие отечественной науки, многолетнюю плодотворную деятельность и в связи с 300-летием со дня основания Российской академии наук (2024 г.).

Уход Рафаэля Мидхатовича – невосполнимая утрата для всех нас.

Сотрудники Санкт-Петербургского Федерального исследовательского центра Российской академии наук, ученики и коллеги, редакционная коллегия журнала Информатика и автоматизация.

Руководство для авторов

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Информатика и автоматизация» http://ia.spcras.ru/. При регистрации авторам рекомендуется заполнить все предложенные поля данных. Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше или LaTex. Объем основного текста (до раздела Литература) - от 20 до 30 страниц включительно. Переносы запрещены. Номера страниц не проставляются. Основная часть текста статьи разбивается на разделы, среди которых являются обязательными: введение, хотя бы один «содержательный» раздел и заключение. Допускается также мотивированное содержанием и структурой материал а выделение подразделов. В основную часть опускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

Author guidelines

Interaction between each potential author and the Editorial board is realized through the pesoal account on the website of the journal “Informatics and Automation” http://ia.spcras.ru/. At the registration the authors are requested to fill out all data fields in the proposed form. The submissions should be prepared using MS Word 2007, LaTex. The text of the paper in the main part should not exceed 30 pages. Pages are not numbered; hyphenations are not allowed. Certain figures, tables, listings and formulas are allowed in the main section, and their typography is considered in more detail at the journal web.

Signed to print 07.11.2024. Passed for print 08.11.2024.
Printed in Publishing center GUAP.
Address: 67 litera A, B. Morskaya, St. Petersburg, 190000, Russia

Founder and Publisher: SPC RAS.
Address: 39 litera A, 14th Line V.O., St.Peterburg, 199178, Russia.
The journal is registered in the Federal Service for Supervision of Communications, Information Technology, and Mass Media,
Registration Certificate (registration number) ПИ № ФС77-79228 dated September 25, 2020
Subscription Index П5513, Russian Post Catalog

Подписано к печати 07.11.2024. Дата выхода в свет 08.11.2024.
Формат 60×90 1/16. Усл. печ. л. 12,26. Заказ № 293. Тираж 300 экз., цена свободная.
Отпечатано в Редакционно-издательском центре ГУАП.
Адрес типографии: Б. Морская, д. 67, лит. А, г. Санкт-Петербург, 190000, Россия

Учредитель и издатель: СПб ФИЦ РАН.
Адрес учредителя и издателя: 14-я линия В.О., д. 39, лит. А, г. Санкт-Петербург, 199178, Россия
Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций,
свидетельство о регистрации (регистрационный номер) ПИ № ФС77-79228 от 25 сентября 2020 г.
Подписной индекс П5513 по каталогу «Почта России»