DOI 10.15622/ia.24.6.8

I. KOTENKO, I. SAENKO, O. LAUTA, V. SADOVNIKOV, E. ICHETOVKIN, W. LI

ANALYSIS OF MODERN RESEARCH ON PROTECTION AGAINST ADVERSARIAL ATTACKS IN ENERGY SYSTEMS

Kotenko I., Saenko I., Lauta O., Sadovnikov V., Ichetovkin E., Li W. **Analysis of Modern Research on Protection against Adversarial Attacks in Energy Systems.**

Abstract. Machine learning-based systems, or machine learning systems, are currently attractive targets for attackers, since disruption of such systems can cause crucial consequences for critical infrastructure, in particular, energy systems. Therefore, the number of different types of cyber attacks against machine learning systems, which are called adversarial attacks, is continuously increasing, and these attacks are the subject of study for many researchers. Accordingly, many publications devoted to reviews of adversarial attacks and defense methods against them appear every year. Many types of adversarial attacks and defense methods in these review articles overlap. However, more recent studies contain information about new types of attacks and defense methods. The purpose of this article is to analyze the research conducted over the past six years in highly ranked journals, with an emphasis on review papers. The result of the study is a refined classification of adversarial attacks, characteristics of the most common attacks, as well as a refined classification and characteristics of defense methods against these attacks. The analysis focuses on adversarial attacks that target energy systems. The article concludes with a discussion of the advantages and disadvantages of various adversarial defense methods.

Keywords: cyber attacks, artificial intelligence, machine learning, adversarial attacks, threat model, defense methods, overview, energy systems, classification.

1. Introduction. Artificial intelligence (AI) and machine learning (ML) are areas of computer science that are increasingly used to create decision support systems that can process and analyze information in the same way as humans do. The key idea behind the operation of ML systems is that they are able to learn on their own by analyzing large amounts of empirical data and then use the acquired knowledge to make decisions in new situations. The effectiveness of ML systems is due to their ability to find hidden patterns in data describing various processes or phenomena much faster than humans. Due to this, ML systems are becoming increasingly widespread in many areas of human activity [1]. However, it should not be forgotten that this field, despite a large number of studies, still remains insufficiently studied, and the number of papers in this area is currently growing rapidly [2]. In view of this fact, it is logical to conclude that an equally young area of informatics and computer science is information security of applied artificial intelligence (AI) models [3, 4]. AI is gradually being integrated into business structures and widely implemented in specific applications [5], a trend which undoubtedly presents new challenges for ML models developers [6].

These trends are particularly relevant to energy infrastructure, which is one of the most critical sectors in modern society [7]. The security of energy infrastructure is essential for the well-being of society. Faults in power grids can be caused by various factors, such as, for example, natural phenomena (lightning, floods), contact with the external environment (trees, birds, animals), wear and tear or aging of equipment. All of which can lead to cascading effects that may harm the economy and safety [8]. Detecting and classifying such faults with high accuracy are crucial for the safety of society. For this reason, ML-based methods and approaches are currently used for their detection and classification. ML models, especially those based on deep learning (DL), are widely used in energy system infrastructure due to the huge volumes of data covering energy networks [9].

Simultaneously, smart grids have become prime targets for intruders [10]. The complexity of ML models and methods introduces numerous vulnerabilities which can be used to implement attacks against these models and methods (adversarial attacks). The first cases of implementation of gradient-based adversarial attacks were demonstrated in [11, 12]. These attacks can violate the confidentiality, integrity, and/or availability of smart grids (SG) [13, 14]. Adversarial attacks are implemented using adversarial examples, which are subtle but non-random perturbations designed to force the ML model to produce incorrect output, such as an incorrect classification of an input sample containing data on the state of the power grid [15]. It should be noted that adversarial attacks can be performed with low computational costs. This is especially true for black-box attacks, which do not require internal information about the operation of the ML system. At the same time, defense methods, even against black box attacks, may require different amounts of computing resources, which depends on the type and object of the attack.

Figure 1 shows a general scenario in which a hacker attacks an ML system that includes an intelligent power plant control system (in terms of ML models) and datasets containing parameters of the state of the energy infrastructure that are used for ML. A hacker performs data poisoning and parameter tampering attacks on datasets used to train ML models (neural networks), as well as evasion attacks and various types of adversarial machine learning (AML) attacks that are aimed at changing the operating parameters of ML models.

The hacker's goal is to force the ML model used in the fault classification system of the Power Plant Management System (PPMS) to incorrectly classify the input sample into an erroneous class. To do this, the hacker changes the label of the target class in the training dataset or makes

some other seemingly insignificant changes to the dataset or ML model. As a result, PPMS begins to misclassify faults, which in turn leads to erroneous identification of their causes and erroneous prediction of consequences. Thus, adversarial attacks can cause potentially catastrophic damage to energy infrastructure if they are not detected in time and adequate protective measures are not taken against these attacks [16].

The number of papers addressing adversarial attacks and defense methods against them grows considerably each year.

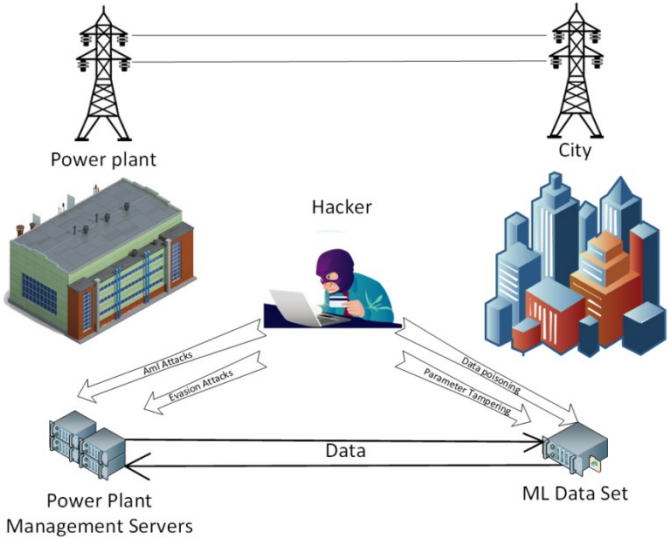

Fig. 1. Impact of adversarial attacks on ML system of energy infrastructure

At the same time, review papers appear every year that clarify the algorithms for implementing such attacks and consider new solutions to counter these attacks. The authors of this paper also proposed approaches and solutions for protecting against attacks on ML systems [17, 18], where an analysis of the existing situation in this area formed the basis of the proposed solutions.

Additionally, it is undoubtedly valuable to conduct a more in-depth analysis of known attacks and defense methods in ML systems. For this purpose, the analysis focused on review papers on adversarial attacks and methods of defense against them that have appeared in recent years. In total, over 30 review papers were selected for analysis for the period 2019-2024. In addition, the papers on adversarial attacks in the field of energy infrastructure for the period 2020-2025 were selected and processed. By

analyzing these papers, one can notice trends in the development of adversarial attacks and defense methods which are very important to know for efficient information protection of the infrastructure of energy systems.

The purpose of this paper is to consider attacks on ML systems and methods of defense against them, based on the analysis of reviews devoted to adversarial attacks as well as research papers devoted to the development and evaluation of attacks and defense methods in the field of energy infrastructure. Therefore, this analysis aims to provide a comprehensive overview of the protection of power systems against attacks on ML systems.

The novelty of the work consists in the synthesis of the latest achievements in the field of defense against attacks on ML systems within one work. At the same time, generalized classification systems for attacks and defense methods are proposed, taking into account known solutions for such classification. In addition, the article provides a description of the most popular adversarial attacks and defense methods. The consideration of the most popular defense methods is performed at a formalized level. Thus, this article serves as a consolidated resource and a practical guide for researchers in the field of information security of ML systems.

The article has the following further structure. Section 2 provides a general description of related papers that were analyzed. Section 3 is devoted to the consideration of the most popular adversarial attacks. Section 4 considers methods of defense against adversarial attacks. The discussion is held in Section 5. Section 6 contains the final conclusions.

2. Analysis of related papers. The analysis of related articles was carried out for two purposes. The first purpose was to familiarize with review papers on the topic of adversarial attacks and protection methods. These studies, as a rule, provided a classification of adversarial attacks and protection methods, and also characterized them. In this case, various areas of application for ML systems affected by adversarial attacks were considered. It should be noted that the classification of adversarial attacks and protection methods in different review papers was different. The second purpose was an attempt to generalize the articles that considered the results of implementing adversarial attacks and protection in the energy field. Here, it was of interest to identify the most popular approaches to performing adversarial attacks and protection methods that were implemented in various areas of managing and monitoring energy networks and their elements.

The selection of review papers for analysis was carried out as follows. The search engines Google.com and Scholar.google.com with the term “Adversarial attack defense review” were used. The search period was set as 2020–2025. As a rule, articles published in journals with quartiles Q1

and Q2 in the index database Scopus were subject to selection (since a large number of papers have been published on the topic, and the high quartile is an indicator of a high level of publication). The quartile was determined based on the data presented in the system https://www.scimagojr.com. A total of 30 review articles were selected for analysis in this way. Table 1 provides general information about the selected review articles.

Table 1. General information about the selected review articles

P_1	P_2	P_3	P_4	P_5	P_6	
Huang X. et al. [19]	2020	Q1	203	11	0.05	Deep Learning
Martins N. et al. [20]	2020	Q1	72	18	0.25	Malicious Code
Oseni A. et al. [21]	2020	Q2	139	25	0.18	Machine Learning
Xu H. et al. [22]	2020	Q2	136	12	0.09	Images, Graphics, Text
Ren K. et al. [23]	2020	Q3	115	4	0.03	Deep Learning
Zhou X. et al. [24]	2020	Q3	13	4	0.31	Electricity, Smart Grids
Akhtar N. et al. [25]	2021	Q1	450	339	0.75	Computer Vision
Zhang H. et al. [26]	2021	Q1	152	34	0.22	Electricity, Smart Grids
Chakraborty A. et al. [27]	2021	Q1	78	1	0.01	Deep Learning
Rosenberg I. et al. [28]	2021	Q2	163	38	0.23	Cybersecurity
Tian J. et al. [29]	2021	Q2	65	42	0.65	Electricity, Smart Grids
Kong Z. et al. [30]	2021	Q3	132	75	0.57	Images, Text, Malicious Code
Zhou Sh. et al. [31]	2022	Q1	185	121	0.65	Deep Learning
Khamaiseh S.Y. et al. [32]	2022	Q1	128	45	0.35	Deep Learning
Liang H. et al. [33]	2022	Q1	52	21	0.40	Deep Learning
Tian Q. et al. [34]	2022	Q1	46	26	0.57	Digital Signals
Anastasiou Th. et al. [35]	2022	Q1	34	27	0.79	Deep Learning
Li Y. et al. [36]	2022	Q2	103	48	0.47	Deep Learning
Tian J. et al. [37]	2022	Q2	49	21	0.43	Electricity, Smart Grids
Li H. et al. [38]	2022	Q2	46	7	0.15	Images
Girdhar M. et al. [39]	2023	Q1	246	140	0.57	Autonomous Vehicles
Goyal Sh. et al. [40]	2023	Q1	176	119	0.68	Text
Al-Khassawneh Y.A. [41]	2023	Q1	53	23	0.43	Cybersecurity
He K. et al. [42]	2023	Q1	166	64	0.39	Cybersecurity
Sun L. et al. [43]	2023	Q1	254	134	0.53	Autonomous Vehicles
Qureshi A.U.H. et al. [44]	2023	Q2	179	48	0.27	Graphics
Hoang V.-T. et al. [45]	2024	Q1	173	136	0.78	Deep Learning
Baniecki H. et al. [46]	2024	Q1	179	131	0.73	Deep Learning
Andrade E. de O. et al.[47]	2024	Q1	65	54	0.83	Images
Zhang Ch. Et al. [48]	2025	Q1	130	122	0.93	Images, Text

Table 1 presents the following indicators: P1 – reference; P2 – year of publication; P3 – quartile according to the Scopus database; P4 – total number of references in the article; P5 – number of references to sources published since 2019; P6 – ratio of the share of references to sources

published since 2019 to the total number of references (i.e., P6 = P5 / P4); P7 – subject area.

Analyzing the data on the P3 indicator, we can conclude that review papers on the topic of attacks on ML systems are of great importance and are of undoubted interest. The vast majority of the analyzed papers were published in highly ranked journals of the Q1 and Q2 quartiles. Only two analyzed papers of 2020, [23] and [24], belong to the Q3 quartile.

The P4, P5, and P6 indicators characterize the power of the reviews. In journals of the Q1 and Q2 quartiles, as a rule, the P4 indicator takes a value greater than 100. The P5 and P6 indicators serve to characterize the relevance of the reviews. For papers published in 2020, the P5 index has a small value (not exceeding 20) which is expected, since the threshold for the P5 assessment was set at 2019. This specific cutoff year was chosen for the purpose of establishing a consistent six-year benchmark for analysis. Publications from 2019 and later represent a six-year window of recent research, which we classify as "fresh". For 2021–2025, the P5 index increases, taking its highest value in 2025. The P6 index shows the degree of "freshness" of the entire review work. As a rule, with an increase in the year of publication of the article, the value of P6 also increases, especially for papers of the Q1 quartile, for which the total number of references P4 exceeds 100. The growth of the values of P5 and P6 with an increase in the year of publication of the article also indicates that, as a rule, in different review papers one can find references to the same original studies, which consider specific cases of the implementation of various types of adversarial attacks and defense methods.

The analysis of the P6 indicator allows us to conclude that most reviews have a broad subject area, which is marked as Deep Learning or Machine Learning (11 reviews). At the same time, there are reviews devoted to a specific subject area. The following subject areas are considered: Images, Graphics, Text and/or Malicious Code – 8 articles; Electricity, Smart Grids – 4; Cybersecurity – 3; Autonomous Vehicles – 2; Computer Vision – 1; Digital Signals – 1. Such a wide range of subject areas in the review papers suggests that the topic of adversarial attacks and defense against them currently affects almost all modern intelligent systems.

Each of the review papers considered the classification of attacks and defense measures, and also provided characteristics and examples of their implementation, taken from originals. It should be noted that the classification of attacks and defense measures was updated from year to year.

Information on the studies that consider the implementation of adversarial attacks and defense measures in the energy field is provided in Tables 2 and 3.

Table 2. Studies on adversarial attacks and defenses in smart grids

Studies	Description of attacks	Description of defense measures	Scope of application
[49]	Adversarial evasion and poisoning attacks	Typical countermeasures against adversarial attacks	Smart grids
[50]	The main attack surfaces and a threat analysis	-	Smart grids
[51]	Different levels of adversarial attacks	Industrial Internet of Things (IoT) architecture and a convolutional neural network with a strategy based on continuous wavelet transform	Demand side management in a smart grid
[52]	Generative Adversarial Network (GAN)	A new zero-sum game based on the Generative Adversarial Networks	Energy System Operators
[53, 54]	Evasion attacks by injecting adversarial samples	An anomaly detector combining an autoencoder, a convolutional-recurrent, and a feed-forward neural networks.	Electricity theft detectors
[55]	Evasion attacks: Distillation, No-Adversarial-Sample-Training, and False-Labeling	Federated learning (FL) and Adversarial training	Smart meters
[56]	Adversarial perturbation, backdoor injection, DoS attacks	A fine-tuned Deep Neural Network (DNN) model	Energy usage control systems
[57]	Universal adversarial attack to deceive deep learning (DL) models	Advanced DL models	Power quality disturbances
[58]	Adversarial attack injection with various reality-imitating techniques	DL monitoring method based on robust feature engineering	Phasor measurement units
[59]	Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini and Wagner attacks	An artificial neural network to analyze a dataset of IEC 60870-5-104 traffic data	The communication network of SGs
[60]	DoS attacks, backdoor injections, and adversarial perturbations	Integration of ML and DL techniques	Smart grids

Continuation of the Table 2

Studies	Description of attacks	Description of defense measures	Scope of application
[61, 62]	Novel attack method called Ensemble and Transfer Adversarial Attack	DL networks	Demand Response mechanisms
[63]	False data injection	-	Smart grids
[64]	Adversarial attacks with distribution	Deep reinforcement learning with a variational auto-encoder, which constrains the deviation of malicious samples	Dynamic pricing system
[65]	Misleading the prediction of deep neural networks	Deep neural networks	Non-intrusive load monitoring
[66]	Adversarial machine learning attacks	DL models	Distributed demand response
[67]	Deep black box adversarial attacks	Quantum voting ensemble models	Smart Power Grids
[68]	Generating adversarial attacks based on linear optimization	Monitoring the observable variables	Electricity markets
[69]	Different cyber adversaries	Different detection techniques, equipment protection plans, and mitigation strategies	Electricity generation, transmission, and distribution sectors
[70]	A novel adversarial attack GAN-GRID	Security mechanisms against adversarial manipulation	The stability prediction system of a smart grid
[71]	Adversarial false data injection evasion attacks	Unsupervised spatial-temporal graph autoencoder	Data-driven detectors
[72]	Inference and evasion attacks	GAN and an ML classifier	Vehicle-to-Microgrid
[73]	Targeted adversarial false data injection strategy	DL models	Bad data detectors
[74]	Complex white-box adversarial attacks	Bayesian Neural Network framework	Smart grid fault prediction
[75, 76]	Overview of attack methods on ML-based smart grids	Overview of defense methods in ML-based smart grids	IoT-based smart grid
[77]	Different adversarial attack scenarios	Mitigation strategy for adversarial attacks based on causal theory	Measurement data protection

Table 3. Studies on adversarial attacks and defenses in energy systems in general

Studies	Description of attacks	Description of defense measures	Scope of application
[78]	Adversarial learning attack	DL models	Renewable energy forecasting
[79]	Targeted, semi-targeted, and untargeted adversarial attacks	Adversarial training	Wind power forecasting
[80]	White-box attacks and black-box attacks	DL models	Solar power generation forecasting
[81]	Targeted universal adversarial attacks	Adversarial training, defensive distillation, and feature squeezing	Power quality disturbance classification
[82]	A novel attack in Reinforcement learning (RL)	A "robustification" of RL algorithms	Prosumer pricing aggregations
[83]	FGSM, FGSM-variant, PGD, and Stochastic Serial Attacks	Multi-Source Feature Detector to detect adversarial attacks	Classification of power quality
[84]	Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) and the Jacobian-based Saliency Map Attack (JSMA)	DL models	Power System State Estimation
[85]	Civil (or Sybil) and Fast Gradient Sign Method attacks	Random Forest, Extreme Gradient Boosting, Decision Tree, Support Vector Machine, and k-Nearest Neighbors.	Wind power forecasting
[86]	Adversarial algorithm attacks	Investing in an energy-storage system	Power system load forecasting
[87]	Synthetic minority class samples	Generative-adversarial-based semi-supervised learning framework	Diagnostics of attacks and faults in electrical networks
[88]	Various attacking objectives	Bayesian training method	Electric load forecasting
[89]	Momentum Iterative FGSM	Iterative adversarial training	Power quality disturbances classification
[90]	Various naturally perturbed data	Physics-constrained adversarial training	Power grids
[91]	Hidden FGSM attacks	ML models	Power distribution systems
[92]	False data injection attacks with adversarial examples	DL models	Bad data detectors
[93]	FGSM and Zeroth Order Optimization	Industrial IoT platform combined with an ensemble ML strategy	Power transformer

Continuation of the Table 3

Studies	Description of attacks	Description of defense measures	Scope of application
[94]	Manipulating Input Data in DNN	DL models	Base station power distribution
[95 – 97]	FGSM, Momentum Iterative FGSM, PGD attacks	DL models	Power allocation
[98]	Data poisoning techniques: scaling and random noise effects	FL models	Energy Forecasting
[99]	Different adversarial attacks	Reactive power injection	Distributed energy resources
[100]	Adversarial ML attacks	ML models	Solar photovoltaics
[101]	False data injection attacks	An unsupervised adversarial autoencoder	Power distribution grids
[102]	The massive interaction of data inside and outside the demand response management system	The adversarial multi-agent reinforcement learning framework	The demand response management system
[103]	Manipulating distributed energy resources	Mitigation strategies to thwart adversaries	Power distribution grids
[104]	Fast Gradient Sign Method	One-class Support Vector Machine and Local Outlier Factor	Photovoltaic Power Generation Forecasting
[105]	Adversarial ML algorithm powered by black-box optimization	DL models	Demand side management
[106]	An evasive deep black-box adversarial attack	ML models	Cyber–Physical Power Systems

Table 2 presents information related to smart grids. The term “Adversarial attacks defense smart grid” was used to select these studies. Table 3 presents information related to energy systems in general. In this

case, the term “Adversarial attacks defense power energy” was used. As a rule, papers published in the period 2020-2025 were subject to selection.

The characteristics of the articles are carried out according to the following criteria: (1) description of the attacks under consideration; (2) description of the protection methods under consideration; (3) the area or element in the energy system that was affected by the attack.

Analyzing the data presented in Tables 2 and 3, we can draw the following conclusions.

1. Adversarial attacks in energy infrastructures can be aimed at any elements or tasks in these infrastructures. For energy infrastructure elements, the targets of adversarial attacks can be both monitoring and state control tools (smart meters, theft detectors, quality violation detectors, bad data detectors, etc.) and elements related to trade and marketing (demand response and demand management mechanisms, pricing, electricity markets, etc.). For the tasks to be solved, the targets of adversarial attacks specified in the studies considered above are most often forecasting tasks (stability and failures of the smart grid, wind energy, solar energy, photovoltaic energy, electrical load, etc.).

The next most common task is the classification of power quality violations. This fact is explained by the fact that forecasting and classification problems in energy infrastructures are solved using various ML models (shallow, deep, federated, reinforcement), the stability of which is primarily targeted by adversarial attacks.

2. The papers considered above examine both new types of adversarial attacks (Ensemble and Transfer Adversarial Attack, GAN-GRID, etc.) and well-known ones (FGSM, GAN, PGD, Carlini & Wagner attack, False Data Injection, etc.). New types of attacks are proposed in order to draw the attention of developers and researchers to the resilience of energy infrastructure to adversarial attacks and encourage them to develop new protection measures. Known attacks were considered in order to demonstrate the capabilities of the proposed approaches to protection.

3. Among the protection methods, the most popular are measures based on the use of ML and DL models. At the same time, there are also original approaches, such as defensive distillation, feature squeezing, continuous wavelet transforms, quantum voting ensemble models, physics-constrained adversarial training and others.

In the next two sections, we consider the most common types of adversarial attacks and defense methods against them.

3. Attacks against artificial intelligence (machine learning) systems. To classify attacks against ML systems, various features are used in review papers [19 – 48], in particular, the following:

– data area under attack (images, audio, text);

– ML system architecture level where the attack is implemented (input, output, generation, gradient, features, application, authentication, database, memory);

– feature characterizing the attack (data redirection, inappropriate data input, data corruption, weak authentication, overflow);

– attack type (model extraction, poisoning, evasion, special type, digital, physical);

– attack method (attacks on parts of the ML model, direct attacks on data);

– attack target (model compromise, evade of detection, inappropriate data input);

– input data type (continuous, discrete);

– knowledge of the attacked system (“white box”, “black box”, “gray box”).

The meanings of some features (data area under attack, feature characterizing the attack, attack method, and input data type) are obvious. However, the meanings of other features require explanation.

The main types of adversarial attacks are model extraction, data poisoning, and evasion. Model extraction is an attacker's analysis of an ML system in order to extract the data on which it was trained [107]. The damage from such an attack is that the training data or the model itself may be confidential. Data poisoning attacks are one of the main security threats at the training stage of ML models. They compromise the integrity of the model by polluting the training data. As a result, the resulting model may be distorted or even unsuitable for further use [108]. Evasion attacks are a special type of adversarial attacks in which malicious data is delivered to the ML system without being detected by conventional security systems [109]. Attacks of this type are particularly widespread in Smart Power Grids, where they can generate adversarial evasion samples by altering malicious consumption data, tricking detectors into classifying them as benign [110].

ML models are increasingly used in security systems, such as Network Intrusion Detection Systems (NIDS). Therefore, attackers set goals for adversarial attacks related to compromising ML models, evading detection, or introducing unwanted data [111]. Compromising the model leads to a decrease in confidence in the results shown by the NIDS and errors in detecting network attacks. Evading detection or introducing unwanted data allows, as mentioned above, one to introduce malicious or unwanted data into the ML system, respectively.

The feature of “knowledge of the attacked system” is used in attack classifications proposed by almost all review papers. The “white box” value

means that the attacker has full knowledge of both the ML model and the training datasets. The “black box” is the opposite case, when the attacker knows nothing about either the ML model or the datasets. An intermediate option is the "gray box", when the attacker has partial, incomplete knowledge of the ML model and data sets.

Tables 4, 5, and 6 present the distribution of the most well-known adversarial attacks by various classification features, taking into account whether the attack belongs to the “black”, “white”, or “gray” box, respectively. The full name of the attack, and its abbreviation, a brief description of the attack, the parameters used by the attack, and the values of the classification features of the attack are given.

The last column shows the *S* indicator, which characterizes the popularity of the attack. The *S* value for each attack is calculated using the following formula:

$$S = \sum_{i=1}^{n} A_i, \tag{1}$$

where A_i is a parameter that shows how many times the value of the *i*-th feature, that the attack has, occurs in Tables 4, 5, and 6.

For example, for the BA attack, the value of A_1 (feature of “Data”) is 18, A_2 (“Level”) is 4, A_3 (“Feature”) is 3, A_4 (“Type”) is 3, A_5 (“Method”) is 12, A_6 (“Target”) is 19, A_7 (“Input type”) is 18. Therefore, for the BA attack, the value of the *S* indicator is 77.

Note that since the total number of attacks presented in Tables 4, 5, and 6 is 31, and the number of classification features is 7, the maximum possible value of S is 217. However, this value is not achievable in practice. The maximum value of *S* is the “black box” attack OPA-98. The minimum value of $S = 40$ is the “gray box” attack BOA.

Thus, each attack has an S indicator, which shows how often the classification features corresponding to this attack are encountered. Consequently, it is possible to distribute the list of attacks presented in the tables into groups, including attacks with the most popular features, where the S value is high (greater than 80), attacks with specific features, where S is less than 70, and attacks with intermediate values. In this way, we can identify groups of attacks, in each of which the set of features will be specific or more general. This can help in designing methods of protection against such attacks, since by identifying a group by features, we can design protection against the entire group at once. And if the S indicator is small, then an individual approach to protection is needed for such an attack.

Table 4. Distribution of “black box” attacks by classification features

Name	Description	Parameters	Data	Level	Feature	Type	Method	Target	Input type	S
Boundary Attack (BA)	Finding class boundaries by changing model output values slightly	Step of changing the output value and the number of iterations	Image	Output	Redirection	Extraction	Direct attacks on data	Model compromise	Continuous	77
Zero-Query Attacks (ZQA)	Minimal input data changes without access to the model	Number of model queries and size of changes	Image	Output	Redirection	Extraction	Direct attacks on data	Model compromise	Continuous	77
Generative Adversarial Network (GAN)	Generating new examples using GAN to create attack samples	GAN architecture and number of training epochs	Image	Generation	Inappropriate data input	Special	Attacks on parts of the ML model	Inappropriate data input	Continuous	69
One Pixel Attack (OPA)	Changing a single pixel in an image to achieve a significant impact on the model's output	Pixel selection and change value	Image	Input	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	98
Zeroth Order Optimization (ZOO)	Estimating the loss function without computing gradients using stochastic methods	Step size and number of loss function evaluations	Image	Output	Redirection	Extraction	Attacks on parts of the ML model	Model compromise	Continuous	73
Natural Evolution Strategies (NES)	Evolutionary algorithms for finding optimal changes to input data	Population size and number of generations	Image	Generation	Inappropriate data input	Special	Attacks on parts of the ML model	Inappropriate data input	Continuous	69
Genetic Algorithms (GA)	Genetic operations for creating new audio samples to introduce unwanted changes	Population size, mutation and crossover probability	Audio	Generation	Inappropriate data input	Special	Attacks on parts of the ML model	Inappropriate data input	Discrete	50
Improved Genetic Algorithm (IGA)	Improved genetic operations for optimizing audio samples	Parameters are similar to GA, but with improved selection and mutation	Audio	Generation	Inappropriate data input	Special	Attacks on parts of the ML model	Inappropriate data input	Discrete	50
Real-World Noise (RWN)	Injecting real noise into audio data to manipulate model output	Noise level and type of sound source	Audio	Input	Data corruption	Poisoning	Direct attacks on data	Model compromise	Continuous	77
Probability Weighted Word Saliency PWWS)	Using probabilistic word weighting to generate attack texts	Word probabilities and their impact on the model	Text	Generation	Data corruption	Special	Direct attacks on data	Model compromise	Continuous	83
Greedy Search Algorithm (GSA)	A greedy search algorithm to find the weakest points in text data	Search strategy and criteria for selecting words	Text	Generation	Data corruption	Special	Direct attacks on data	Model compromise	Continuous	83
Insertion and Removal of Words (IRW)	Modifying text by adding or removing words to manipulate	Number of words inserted/removed and their context	Text	Generation	Data corruption	Special	Direct attacks on data	Model compromise	Continuous	83

Table 5. Distribution of “white box” attacks by classification features

Name	Description	Parameters	Data	Level	Feature	Type	Method	Target	Input type	S
Fast Gradient Sign Method (FGSM)	Modify pixels based on the gradient of the model loss	The parameter ε determines the rate of change	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Iterative Gradient Sign Method (IGSM)	Iteratively modify pixels, accumulating changes for stronger attacks	The number of iterations and the step of change	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Jacobian Saliency Map Attack (JSMA)	Modify pixels using the importance of gradients to select the most vulnerable pixels	Determining the importance of each pixel	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Block Input Manipulation (BIM)	Modify blocks of pixels using gradients to fine-tune the changes	The block size and the number of iterations	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Undetectable Perturbation (UP)	Invisible changes that are not noticeable to humans but affect the model output	Special algorithms for minimizing distortions ns	Image	Input	Data corruption	Evasion	Direct attacks on data	Evade of detection	Continuous	85
Feature Adversary (FA)	Changing features based on importance analysis to confuse the model without changing the input data	Definition of important features of the model	Image	Features	Data corruption	Evasion	Attacks on parts of the ML model	Evade of detection	Continuous	73
Carlini and Wagner's Attack (C&W)	Optimized changes that minimize the distance between the original and modified images	Weighting factors for different targets	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Iterative Least-Likely Class Method (ILLCM)	Minimizing the probability of the correct class through iterative changes to the input data	Number of iterations and target class	Image	Gradient	Data corruption	Evasion	Direct attacks on data	Evade of detection	Continuous	86
One-Step Target Class Method (OSTCM)	Directed changes that cause the model to misclassify data into a predetermined class	Defining the target class and changing the input data	Image	Output	Data corruption	Evasion	Direct attacks on data	Evade of detection	Continuous	83
Deep Fool (DF)	Iteratively changing the input data to minimize the distance to the class boundary	Number of iterations and change step	Image	Gradient	Data corruption	Digital	Direct attacks on data	Model compromise	Continuous	97
Hot/Cold method (HCM)	Temperature changes in the input data that affect the model's classification process	Temperature and exposure time	Image	Input	Data corruption	Evasion	Direct attacks on data	Evade of detection	Continuous	85
Ground-Truth Attack (GTA)	Incorrect class labels fed to the model to manipulate its output	Type of mislabeled label and its impact on learning	Image	Input	Data corruption	Evasion	Attacks on parts of the ML model	Evade of detection	Discrete	78
Targeted Audio Adversarial Examples (TAAE)	Audio adversarial examples aimed at fooling speech recognition systems	Using psychoacoustic principles	Audio	Input	Data corruption	Evasion	Attacks on parts of the ML model	Evade of detection	Continuous	62

Table 6. Distribution of “grey box” attacks by classification features

Name	Description	Parameters	Data	Level	Feature	Type	Method	Target	Input type	S
Cross-Site Scripting (XSS)	Malicious code is introduced through vulnerabilities in web applications, allowing the attacker to execute scripts on behalf of the user	Exploits vulnerabilities in web applications; parameters may include script type and injection location	Text	Application	Inappropriate data input	Poisoning	Direct attacks on data	Model compromise	Discrete	65
Password Guessing (PG)	Password guessing through systematic testing of possible combinations or using dictionaries	Parameters include password length, hashing algorithm complexity, and dictionaries used	Text	Authentication	Weak authentication	Poisoning	Direct attacks on data	Model compromise	Discrete	60
Cross-Site Request Forgery (CSRF)	Forging requests to a web application on behalf of a user without their knowledge	Exploits vulnerabilities in user authentication; parameters may include request type and target page	Text	Application	Inappropriate data input	Poisoning	Direct attacks on data	Model compromise	Discrete	65
SQL Injection (SQLI)	Inserting malicious SQL code into database queries in order to gain access or manipulate data	Parameters may include database type and table structure; used to gain unauthorized access to data	Text	Database	Inappropriate data input	Poisoning	Direct attacks on data	Model compromise	Discrete	64
Buffer Overflow Attack (BOA)	Buffer overflow by writing more information than allocated, which can lead to arbitrary code execution	Exploits memory management vulnerabilities; parameters may include buffer size and vulnerability type	All	Memory	Overflow	Physical	Attacks on parts of the ML model	Evade of detection	Continuous	40
Weak Authentication Attack (WAA)	Exploiting weak passwords or authentication vulnerabilities to gain unauthorized access to systems	Parameters may include password complexity and authentication methods; used to bypass security systems	All	Authentication	Weak authentication	Physical	Direct attacks on data	Model compromise	Discrete	52

As a result, we obtain the following distribution of attacks by popularity groups:

– Group 1 “High Popularity” (S > 80) includes 14 attacks: OPA, PWWS, GSA, IRW, FGSM, IGSM, JSMA, BIM, UP, C&W, ILLCM, OSTCM, DF, HCM;

– Group 2 “Medium Popularity” (70 < S ≤ 80) includes 9 attacks: BA, ZQA, GAN, ZOO, NES, RWN, FA, GTA, TAAE;

– Group 3 “Low Popularity” (S ≤ 70) includes 8 attacks: GA, IGA, XSS, PG, CSRF, SQLI, BOA, WAA.

Thus, we have obtained 3 groups into which the above-mentioned adversarial attacks are divided according to the set of features. Group 1 represents attacks that can be described by features that are often used to describe attacks. This allows using a universal approach to detect them. Attacks from Group 2 contain rarer features. Probably, approaches from group 1 will be partially suitable for their detection. However, these attacks have their own specifics. But to detect attacks from Group 3, separate individual approaches are needed.

The proposed approach to classifying adversarial attacks can become the basis for further research and practical experiments.

Let us consider in more detail the most popular adversarial attacks.

FGSM is a white-box attack against neural networks that is used to fool models trained to recognize images. FGSM involves slightly altering an image so that the trained model will mistakenly identify it as a different class. This is done using the gradient descent method, which allows finding the most sensitive pixels in the image. When using FGSM, the initial image is considered as a point on the path from the original to the altered image that provides the maximum change in the rate value of the target loss function. This can be written as follows [112]:

$$Z^* = Z + \epsilon \cdot \text{sign}\big(\nabla_Z J(\theta, Z, W)\big), \quad (2)$$

where Z^* – adversarial image, Z – original image, ∇_Z – gradient of the image Z (it shows a directional change in the intensity or color in an image), W – label associated with input image, ϵ – applied noise, θ – model's parameters, and J – loss function.

The gradient of the loss is then calculated for each pixel in the image, after which all pixels with the smallest gradient magnitude are set to zero, and the rest are increased or decreased by an amount that is the sign of the gradient. The FGSM method allows one to create fake images that look almost identical to the originals, but carry altered information that can fool an ML model.

There are several variants of FGSM, which differ in how the gradient descent step size is determined. For example, FGSM can be used with a

fixed step, or it can determine the step at each point using a backtracking line search. FGSM has limitations. One of them is that the method can fool the model only to a certain extent, after which the results are no longer reliable, and the model begins to identify the altered image correctly. In addition, FGSM can only be applied to models that use gradient descent for training. It should also be noted that attack methods such as FGSM can be used not only by attackers, but also for various research tasks related to assessing the level of defense of neural networks and their behavior in various scenarios. In particular, the FGSM method can be used to develop new algorithms for protecting neural networks, allowing to increase the level of defense against such attacks.

IGSM is a variation of the white-box attack method against neural networks that extends the capabilities of a similar FGSM algorithm. It is based on repeated applications of the FGSM method with several modifications [113]:

$$Z'_0 = Z;\ Z'_{n+1} = Clip_{Z,\epsilon}\{Z'_n + \alpha \cdot \text{sign}(\nabla_Z J(\theta, Z'_n, W))\}, \quad (3)$$

where α – step length, $Clip_{Z,\epsilon}\{A\}$ – the element-wise clipping of Z.

IGSM is an optimization algorithm that starts with an original image and continues updating it through a series of iterations using FGSM. In each iteration, the pixel values are changed in the direction of increasing the loss of the target function. Unlike FGSM, which uses only one iteration to generate fake images, IGSM repeats the attack procedure in each iteration, which gives a better effect but requires more computational resources. IGSM can be used for both targeted and non-targeted attacks.

JSMA is a white-box attack algorithm for image counterfeiting detection systems based on deep learning methods. This algorithm computes the sensitivity within the gradient of the loss function (L) for the input image (x), leading to favorable outcomes in white-box target-specific attacks [114]. Their execution flow is as follows: $\triangle L_x = \partial L_t / \partial x_i$. Then, the adversarial samples are calculated based on the algorithm's forward derivatives by the Jacobi matrix, that is computed by

$$\varphi(x,t)[i] = \begin{cases} 0, \ \ if \ \ \dfrac{\partial L_t(x)}{\partial x_i} < 0 \ \ or \ \ \displaystyle\sum_{i \neq t} \frac{\partial L_t(x)}{\partial x_i} > 0, \\ \dfrac{\partial L_t(x)}{\partial x_i} \left| \dfrac{\partial L_t(x)}{\partial x_i} \right|, \ \ \text{otherwise}, \end{cases}, \quad (4)$$

where i represent the input feature. The matrix will compute the change in positions of pixels to obtain the target t in the input image.

As a result, JSMA can identify the most "important" features (parts) of an image that influence the model's classification. JSMA starts by selecting a target model to attack. It then calculates a gradient vector for each part of the image, allowing it to identify those parts that influence the classification of the image as a fake. It then increases the influence of these parts of the image while decreasing the influence of other parts, leading to a result where the neural network classifies the image as a fake.

BIM is based on introducing changes into the input data that is fed to the trained model [115]:

$$f_\theta \to f_{\theta'} \Longrightarrow \begin{cases} \forall x,\ g(f_\theta, x) \neq g(f_\theta, x'), \\ \forall x,\ f(x) \not\approx f(x'), \end{cases} \tag{5}$$

where f_θ is a model with the model's parameters θ; $g(\cdot,\cdot)$ denotes an explanation function, for which both model and data are the input, where the output domain varies between different explanation methods.

The results of an attack can lead to erroneous conclusions or incorrect system actions, which can threaten security and create risks for the business.

Examples of BIM attacks include changing the values of input parameters when training ML models. For example, if a system is trained to identify different patterns based on color, size, and shape, an attacker can trick the system by providing it with input data that contains altered values, such as changed colors or object sizes.

Another example of a BIM attack can be aimed at automated quality control systems, when an attacker sends corrupted data, creating false error signals. Such attacks can cause system failures, equipment malfunctions, or dangerous disruptions in the manufacturing process.

To prevent BIM attacks, it is necessary to include security measures in the design and configuration of AI systems, such as input data validation, data integrity control, and training models on a large amount of data. It is also worth using additional control methods, such as the use of one-time PIN codes or two-factor authentication, in case the system interacts with important data.

UP is a type of white-box attack that involves introducing minor changes to the data or parameters of a model that lead to erroneous conclusions and discredit the results [116]:

$$x \to x' \Longrightarrow \begin{cases} g(f_\theta, x) \neq g(f_\theta, x'), \\ f(x) \approx f(x'). \end{cases} \quad (6)$$

The main goal of such an attack is to bypass the security system and create distorted data so that it is accepted as correct. UP attack can be used in various fields, for example, to manipulate voting results, change weather forecasts, autonomous vehicles, medical diagnostic systems, etc. It is very difficult to detect, as it creates minor changes in the data.

FA is aimed at allowing attackers to penetrate a system that uses ML and change its behavior in order to gain unauthorized access to data or perform other actions to suit their interests. FA attacks can be implemented using various methods, including introducing false data into the system to change its behavior, using black-box attacks on ML models, analyzing the behavior of the system, etc. These attacks can be quite complex. They require specialized skills and knowledge in the field of ML from the attackers [117].

One way to protect against FA attacks is to install software that can detect abnormal behavior of the ML system and block any attempts by the attacker to interfere. It is also necessary to ensure the defense of personal data, as well as control over the correct operation of the ML system. In addition, it is possible to use secure algorithms and technologies to combat built-in vulnerabilities.

BA is a typical black-box attack method based on the classification boundary. Starting from the original adversarial image, it uses binary search to find a sample point that is near the classification boundary. It performs a random walk along the boundary between two opposite regions, thereby decreasing the distance from the target image. According to this iterative method, the distance from the original image is gradually reduced. In other words, the classification result obtained by the input query of the classifier is always the category to be misclassified [118].

The reason this type of algorithm is called a "boundary attack" is that it generates adversarial examples by searching along the boundary until they converge to produce an optimal or rational solution. The results obtained by this method can satisfy the misclassification requirements of a "black box" model. The total perturbation that is increased compared to the original image depends on the performance of the algorithm.

ZQA is a black-box attack that is designed to transfer knowledge between models without access to the input data [119]. They rely on transferring knowledge between models using the outputs of the models rather than the input data. Traditionally, transferring knowledge between models requires access to the original model's input data, which can lead to

the leakage of sensitive information. However, ZQA attacks can transfer knowledge from one model to another using only the model's output data, significantly reducing the risk of information leakage.

Attackers can use ZQA attacks to perform various tasks, such as creating fake images and videos that fool AI-driven computer vision systems or attacking secure facial recognition systems using data obtained from other recognition systems. In addition, ZQA attacks can be used to identify sensitive information. For example, attackers can use them to detect keywords and phrases in documents that should not be publicly available. They can use the knowledge gained from one model to train another model that can identify this sensitive data.

At the same time, ZQA attacks can also be used for constructive purposes. For example, they can be used to transfer experience between models in the field of medicine or scientific research, allowing the model to speed up the training process and create more accurate results.

Defenses against ZQA can be implemented based on anomaly detection and supervised learning methods. These methods allow identifying unusual output data that can be associated with ZQA. Another possible direction of defense is to develop methods for detecting and preventing experience transfer between models using only output data. Additionally, defense can be strengthened by training models, which helps to cope with attacks and slow down the process of transferring experience between models.

NES is a family of numerical optimization algorithms for black-box problems. Like all other evolution strategies, they iteratively update the parameters of the search distribution, following the natural gradient towards higher expected fitness [120]. The general procedure for implementing this method is as follows. A parameterized distribution is used to generate a set of search points. At each point, the fitness function is estimated. The parameters of the distribution allow the algorithm to adaptively fix the values of the fitness function. For example, in the case of a Gaussian distribution, they include the mean and the covariance matrix. Based on the samples, the search gradient towards higher expected fitness is estimated. Then, an ascent step along the natural gradient is performed. This step is crucial, as it prevents oscillations, premature convergence, and undesirable effects arising from the given parameterization. The entire process is repeated until the stopping criterion is met.

GAN is a black-box attack method that uses neural networks to perform various malicious attacks on ML models. The principle of GAN is to train two neural networks, a generator and a discriminator, which are fed data sequentially to each other and learn from each other. In the first stage,

the generator creates fake data examples that are fed to the discriminator along with real examples from the training set [121]. The discriminator learns to distinguish real data from fake, and the generator learns to create data that is difficult to distinguish from real. In the second stage, the generator uses the acquired knowledge about the structure of the data to create malicious attacks on the ML model. These attacks can be different depending on the type of model and the problem it solves. Once the attack is generated, it can be used by an attacker to attack the ML model. Thus, neural GAN allows generating different types of attacks against ML models, which makes them more vulnerable to attacks. This can be used to test the robustness of models and find vulnerabilities in their defenses.

OPA is a black box attack and is based on ML algorithms. It exploits vulnerabilities in neural networks that identify images based on the color values of each pixel [122]. The basic principle of this attack is to change the value of just one pixel in the image so that the neural network incorrectly classifies this image. For example, when editing a photo of a cat, the OPA attack can change the value of the pixel in the place of the cat's nose so that the neural network will believe that it is actually a dog. The OPA attack uses evolutionary algorithms that allow one to determine the optimal pixels and change their values in such a way as to deceive the neural network. Using such algorithms allows one to achieve maximum attack efficiency with a minimum number of changes to the image.

4. Methods of defense against attacks on ML systems. Possible classification features for methods of defense against attacks on ML systems, indicated in review papers [19 – 48], are:

- direction of the defense method;
- implementation complexity;
- applicability level;
- protected data type;
- impact character;
- operating principle.

By direction, defense methods are divided into defensive and counterattacking. Defensive methods are aimed at protecting against specific types of adversarial attacks. Counterattacking methods are aimed at preventing adversarial attacks in general.

By implementation complexity, defense methods are divided into simple and complex. Simple methods are easy to implement and require minimal computing resources. Complex methods require additional calculations, a complex training process, or specialized implementation.

By applicability level, defense methods are divided into universal and specific. Universal methods can be applied to various types of models

and tasks. Specific methods require a specialized approach or are applied only to certain architectures of machine learning systems.

By the protected data type, there are methods that defend input data (images) and methods that protect neural network weights, i.e., defend the parameters of the machine learning model.

By the impact character, defense methods are divided into active and passive. Active methods change the data or model to increase stability. Passive methods analyze data and detect anomalies without changing the data or the model.

By operating principles, methods are divided into those that perform data (image) transformations, model training, and attack detection and filtering. Transformation methods change the input data. For example, they add noise or perform compression. Methods that affect training change the model training process. For example, these include adversarial learning and distillation. Attack detection and filtering methods analyze and filter the input data to detect attacks.

Table 7 displays the classification feature values for the most common adversarial attack protection methods.

The last column of Table 7, by analogy with Tables 4–6, presents the *S* popularity index of the protection method. It is calculated using a formula similar to the popularity of adversarial attacks. The S index for protection methods ranges from 45 (for the MagNet, DeepHunter, LID and PCA methods) to 66 (for the Mixup method). Therefore, we propose dividing all the protection methods into the following two groups by popularity:

- Group 1 “High Popularity” ($S > 55$) includes 8 attacks: Randomized Smoothing, JPEG Compression, Bit Depth Reduction, Total Variance Minimization, Feature Squeezing, Mixup, L2 Regularization, Median Filtering;
- Group 2 “Low Popularity” ($S \leq 55$) includes 9 attacks: Adversarial Training, Defensive Distillation, MagNet, DeepHunter, LID, PCA, Weight Clipping, Dropout, Spectral Normalization.

The methods of the first group have a wider scope of application. They allow detecting and counteracting a larger number of adversarial attacks. At the same time, the implementation of these methods is usually simple.

The methods of the second group are aimed at protecting against a smaller number of attacks. They are distinguished by increased complexity and specificity.

Let us consider in more detail the most popular adversarial attacks.

Table 7. Classification feature values for the most common adversarial attack defense methods

Defense method	Direction	Complexity	Applicability level	Protected data type	Impact character	Operating principle	S
Randomized Smoothing	Defense	Complex	Universal	Input	Active	Transformation	65
JPEG Compression	Defense	Simple	Universal	Input	Active	Transformation	62
Bit Depth Reduction	Defense	Simple	Universal	Input	Active	Transformation	62
Total Variance Minimization	Defense	Complex	Universal	Input	Active	Transformation	65
Adversarial Training	Defense	Complex	Specific	Weights	Active	Training	54
Feature Squeezing	Defense	Simple	Universal	Input	Active	Transformation	62
Defensive Distillation	Defense	Complex	Specific	Weights	Active	Training	54
Mixup	Defense	Complex	Universal	Input	Active	Training	66
MagNet	Counterattack	Complex	Universal	Input	Passive	Detection and filtering	45
DeepHunter	Counterattack	Complex	Universal	Input	Passive	Detection and filtering	45
Local Intrinsic Dimensionality (LID)	Counterattack	Complex	Universal	Input	Passive	Detection and filtering	45
Principal Component Analysis (PCA)	Counterattack	Complex	Universal	Input	Passive	Detection and filtering	45
Weight Clipping	Defense	Simple	Specific	Weights	Active	Training	51
Dropout	Defense	Simple	Specific	Weights	Active	Training	51
L2 Regularization	Defense	Simple	Universal	Weights	Active	Training	58
Spectral Normalization	Defense	Complex	Specific	Weights	Active	Training	54
Median Filtering	Defense	Simple	Universal	Input	Active	Transformation	62

The idea behind the **Randomized Smoothing** method can be expressed by the following formula [123]:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f(x_i + \epsilon_i), \tag{7}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is some noise introduced into the initial data vector $\{x_i\}$, which has a normal distribution law $\mathcal{N}$ with a zero first moment and a second moment equal to σ^2; $f(\cdot)$ is the model training function; $\hat{y}$ is the result of logical inference on the model.

In other words, Randomized Smoothing is a method for protecting against adversarial attacks that relies on adding random noise to the input data and then averaging the model's predictions.

In the formula, $\hat{y}$ is the average prediction of the model after applying noise. Each image x_i is modified by adding random noise ϵ_i taken from a normal distribution $\mathcal{N}(0, \sigma^2)$. In this way, the model receives several noisy versions of the same image and makes predictions for each of them. Averaging these predictions helps reduce the impact of small changes that can be used to create adversarial examples. This method is effective due to its ability to reduce the sensitivity of the model to small changes, making the model more resilient to attacks.

Adding noise from a normal distribution allows the model to account for the uncertainty in the input data and make more stable predictions. The choice of the parameter σ plays a key role. Too small a value of σ may not provide sufficient protection, while too large a value of σ may result in a decrease in the accuracy of the model on normal data. Randomized Smoothing also requires additional computation, since it must make several predictions for each input image [124]. However, its versatility and ease of implementation make it quite popular for improving the resilience of models to adversarial attacks. This method is especially useful in situations where protection is required against small changes that may be difficult to detect with the human eye.

JPEG Compression is a defense method that is based on applying the JPEG compression algorithm to images before feeding them to the model input [125]. The idea of the method can be represented by the following formula:

$$I_{compressed} = \text{JPEG}(I, q), \tag{8}$$

where I is the original image; q is a parameter describing the image quality after compression; $I_{compressed}$ is the compressed image; JPEG (·) is the image compression algorithm.

The formula describes the process of transforming the original image I into a compressed image $I_{compressed}$ with a given compression quality q. JPEG compression removes high-frequency components of the image, making it less sensitive to small changes that can be used to create adversarial examples. This is because such changes are often in the high-frequency region of the spectrum, which is removed by compression. Thus, the model becomes less susceptible to such attacks, since it processes already “averaged” data.

The choice of the compression quality parameter q is an important aspect of this method. Too low a compression quality can lead to significant information loss, which will negatively affect the performance of the model on normal data [126].

At the same time, too high a compression quality may not provide sufficient protection against adversarial attacks. Therefore, it is important to find a balance between compression quality and model robustness. Despite its simplicity, JPEG Compression can be a very effective defense, especially if adversarial attacks exploit small changes that are easily removed by compression.

Bit Depth Reduction is a method that reduces the bit depth of an image's pixels, thereby reducing its sensitivity to small changes [127]. The idea behind the method is represented by the following formula:

$$I_{reduced} = \frac{\lfloor I \cdot 2^b \rfloor}{2^b}, \quad (9)$$

where I is the original image; $I_{reduced}$ is the transformed image; b is the bit depth.

The original image I is transformed into an image with a reduced bit depth b. The rounding operation $\lfloor I \cdot 2^b \rfloor$ replaces pixel values with the nearest integers in the new bit depth, which significantly reduces the number of possible pixel values. This makes the model less susceptible to small changes that can be used to generate adversarial examples, since they are simply lost in the limited range of values [128].

Total Variance Minimization (TVM) is a method aimed at removing small variations from an image by minimizing pixel variation [129]. The method is described by the following formula:

$$TVM = \min_x ||\nabla x||_1 + \lambda ||x - y||_2^2 . \quad (10)$$

The formula describes the optimization process, where $||\nabla x||_1$ is the L1 norm of the image gradient that minimizes pixel variation, and $\lambda ||x - y||_2^2$ is a regularization term that controls the balance between minimizing variation and preserving the original image y.

TVM uses regularization to remove small changes that can be used to generate adversarial examples. This method is especially useful for images containing noise or other unwanted changes.

The regularization coefficient λ plays an important role in this method. It controls how much we strive to preserve the original image y

while minimizing variation. If λ is too large, the model will tend to preserve the original image with minimal changes, which may not provide sufficient protection against adversarial attacks [130]. On the other hand, too small a value of λ can lead to significant blurring of the image, which can also negatively affect the performance of the model. TVM requires solving an optimization problem, which can increase computational costs, but its results can significantly improve the model's resistance to various types of attacks.

Adversarial Training is a method for training a model on a combination of original and adversarial examples [131]. The method uses two neural networks: a generator that generates adversarial examples, and a discriminator that evaluates them [132]. The idea of the method is expressed by the following formula:

$$L(\theta) = E_{(x,y)\sim D}[l(f_\theta(x), y)] + E_{(x',y)\sim D'}[l(f_\theta(x'), y)]. \quad (11)$$

The formula describes a loss function that consists of two parts. The first part corresponds to the loss on the original data *D*, and the second part corresponds to the loss on the adversarial data *D′*. Training a model on adversarial examples makes it more resilient to such attacks, since it has already "seen" and processed such examples during training.

This method requires the creation of adversarial examples, which may require additional computation and time. Adversarial examples are created by making small changes to the original data so that they cause the model to make erroneous predictions. The model is then trained on these examples, which makes it more resilient to such attacks. One drawback of the method is that it can lead to some decrease in the model's performance on ordinary data. However, Adversarial Training remains one of the most effective methods for protecting against specific types of adversarial attacks.

The process of adversarial training can be broken down into several steps:

1. Creating two neural networks: a generator and a discriminator.
2. Training the generator to create fake data, such as images, sounds, texts, etc. At the same time, the discriminator learns to distinguish fake data from real data.
3. The generator and discriminator compete with each other within the framework of the task that was defined for training the model (for example, face recognition).

4. The generator creates new fake data, and the discriminator evaluates how similar it is to real data. The discriminator's assessment is passed back to the generator so that it can improve its skills.

5. The discriminator learns to become more and more accurate in recognizing fakes. The generator, in turn, learns to create better and better fakes. This process is repeated many times until the fakes become almost indistinguishable from real data.

6. The model obtained after training can be used to analyze new data. For example, it can be used to recognize faces in photographs attached to loan applications, to detect lies during a conversation with a client, etc.

Depending on the applications and areas where this method is applied, the steps may vary slightly to achieve better results. However, the general approach is to adversarially train two networks, a generator and a discriminator, to create better fake data and protect ML systems from attacks and hacks.

Feature Squeezing is a method that reduces feature diversity by applying filters or bit compression [133]. The idea of the method is reflected in the following formula:

$$z' = \lfloor z \cdot 2^b \rfloor / 2^b. \quad (12)$$

The formula describes the process of reducing the bit depth of features z. Features are image channel values or other data characteristics that are used by the model for classification. Bit depth compression reduces the number of possible feature values, making the model less sensitive to small changes that can be used to generate adversarial examples. This method is especially useful for models that are sensitive to small changes in features.

Despite its simplicity, Feature Squeezing can be a very effective defense. Reducing the bit depth of features limits the model's ability to exploit small changes to manipulate classification results. However, too aggressive compression can lead to significant information loss, which can negatively affect the model's performance on normal data. Choosing the right bit depth b is a key factor in the success of this method [134]. It is important to find a balance between reducing the bit depth and preserving the information needed for accurate classification.

Defensive Distillation is a method of training a model using distillation to improve its robustness to adversarial attacks [135]. The method is described by the following formula:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}. \quad (13)$$

The formula describes the process of obtaining soft labels p_i from the model outputs. Here, z_i are the logits (outputs before the activation function) of the model, and T is the temperature, which controls the "softness" of the output probabilities. Increasing the temperature makes the probabilities softer, which helps the model become less sensitive to small changes in the input. This method is especially useful for models that are prone to overfitting and sensitive to small changes in the data.

Distillation involves training a model on soft labels obtained from another model that was trained on the original data. Soft labels represent probabilities of class membership. This approach helps the model learn from more general representations of the data, making it more resilient to adversarial attacks. However, choosing the right value of the temperature T can be challenging, as too high a temperature can led to a decrease in the accuracy of the model, while too low a temperature will not provide sufficient protection [136]. Defensive Distillation requires a specialized training process and can be more difficult to implement compared to other methods. However, its results can significantly improve the resilience of the model.

Mixup is a method that creates new training data by linearly mixing two samples and their labels [137]. The idea of the method is shown in the following formula:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j, \quad \tilde{y} = \lambda y_i + (1-\lambda)y_j. \quad (14)$$

The formula involves mixing the input data x_i and x_j and their labels y_i and y_j with weights λ and (1−λ). This method helps the model become more robust to various types of attacks, as it has already "seen" and processed such examples during training. Linear data mixing creates new points in the feature space, which makes the model more robust and less sensitive to small changes.

The Mixup method requires generating new data during training, which can increase the overall training time. However, using it can significantly improve the model's robustness to adversarial attacks, as it learns from more diverse data. The parameter λ plays a key role in this method: it determines the degree to which the two samples are mixed. Choosing the right distribution for λ (e.g., Beta distribution) can help achieve a better result [138]. Despite the additional computational costs,

Mixup remains a popular method for improving the robustness of models to various types of attacks.

MagNet is a method that uses autoencoders to reconstruct original images and detect adversarial examples [139]. The method is described by the following formula:

$$\min_{\theta} E_{x\sim P_{data}}\left[|x - g_\theta(f_\theta(x))|^2\right]. \quad (15)$$

The formula describes the training process of an autoencoder, where θ are the model parameters, $f_\theta(x)$ is the encoder that transforms the input image into a latent representation, and $g_\theta(f_\theta(x))$ is the decoder that transforms the latent representation back into an image. The goal of training is to minimize the mean squared error between the original image *x* and the reconstructed image $g_\theta(f_\theta(x))$. This allows the autoencoder to efficiently reconstruct the original images and detect adversarial examples.

Since adversarial examples often have noticeable differences from normal data, they can be easily detected through large reconstruction errors. The autoencoder is trained on normal data and then used to check the input images for anomalies. This method requires significant computational resources to train the autoencoder, but its effectiveness in detecting adversarial examples makes it a valuable tool for defending artificial intelligence systems [140]. In addition, MagNet can be adapted to different types of data and models, increasing its applicability.

DeepHunter is a tool for automatic detection of adversarial examples using ensembles of models and statistical methods [141]. The idea is represented by the following formula:

$$d(x) = \frac{1}{N}\sum_{i=1}^{N}\left(f_i(x) - \bar{f}(x)\right)^2. \quad (16)$$

The formula describes the process of calculating the standard deviation of the models' predictions. Here, $f_i(x)$ is the prediction of the *i*-th model, and $\bar{f}(x)$ is the average prediction of all models. If the models' predictions diverge greatly, this may indicate the presence of an adversarial example. DeepHunter exploits the idea that different models can interpret the same data differently, and if most models agree on a classification, this indicates a normal image.

Using ensembles of models allows DeepHunter to achieve high accuracy in detecting adversarial examples. This method requires

significant computational resources to train multiple models and analyze their predictions, but its results can significantly improve the system's resilience to attacks. DeepHunter is especially useful in cases where high detection reliability is required, as it uses the combined efforts of multiple models to detect anomalies [142]. This method is passive, as it analyzes the data and detects anomalies without changing the data or the model.

Local Intrinsic Dimensionality (LID) is a method for estimating the local intrinsic dimensionality of data to detect anomalies [143]. The method is described by the following formula:

$$\mathrm{LID}(x) = \lim_{r \to 0} \frac{\log N(x, r)}{\log r}. \quad (17)$$

The formula describes the process of estimating LID for a data point x. Here $N(x, r)$ is the number of points within a radius r of the point x. Local intrinsic dimensionality is estimated by analyzing the distances between neighboring points in the feature space. Adversarial examples often have higher local dimensionality than normal data, which allows this method to effectively detect them.

The LID method requires calculating the distances between all data points, which can be computationally expensive, especially for large data sets. However, its results can significantly improve the resilience of the system to adversarial attacks. LID is a passive method, as it analyzes the data and detects anomalies without changing the data or the model [144]. This method is especially useful in cases where high accuracy and reliability of adversarial example detection is required, as it uses data properties to detect anomalies.

Principal Component Analysis (PCA) is a method of principal component analysis for anomaly detection [145]. The method is described by the following formula:

$$X_{\mathrm{PCA}} = U \Sigma V^{\mathrm{T}}. \quad (18)$$

In the formula, X_{PCA} is the original data, U is the matrix of eigenvectors, Σ is the diagonal matrix of singular values, and V^{T} is the transpose of the matrix of eigenvectors. PCA transforms the data into a principal component space, where each component represents the direction of maximum variance. This allows for anomaly detection, since adversarial examples often reside in areas of high variance.

PCA is a powerful tool for anomaly detection, since it can detect data that deviates significantly from a normal distribution. However, it requires performing singular value decomposition on the data, which can be computationally expensive for large datasets. PCA is a passive method, since it analyzes the data and detects anomalies without changing the data or the model. This method is especially useful in cases where high accuracy and robustness in detecting adversarial examples are required.

Weight Clipping is a method that limits the values of a model's weights within a certain range [146]. The method is described by the following formula:

$$w' = \max(-c, \min(c, w)). \tag{19}$$

The formula describes the process of trimming the weights w to the limits [$-c$, c], where c is the weight limit. This method is based on the idea that too large or too small weight values can make the model more sensitive to small changes in the data. Weight clipping helps control this sensitivity, making the model more resilient to adversarial attacks. This method is especially useful for models that are prone to overfitting and sensitive to large weight values.

Choosing the right value of the limit c is the key to the success of this method. If the limit is too large, it may not provide sufficient protection, while too small a limit can lead to a significant decrease in the model's performance on normal data. Weight clipping is a simple method that is easy to implement and apply to any model. Its advantage is that it does not require significant changes to the model architecture and can be easily integrated into an existing data processing pipeline. This method is especially useful for problems where control over the scale of the model's weights is required.

Dropout is a regularization method based on randomly switching off neurons during training [147]. The method is described by the following formula:

$$h_i = \begin{cases} 0 \; with \; probability \; p, \\ \dfrac{x_i}{1-p} \; otherwise. \end{cases} \tag{20}$$

In the formula, h_i is the output of the neuron after applying the Dropout method, p is the probability of the neuron being disconnected, and x_i is the input of the neuron. This method helps prevent overfitting and

makes the model more resilient to adversarial attacks, as it forces the model to use different paths to classify data.

The Dropout method is a simple and effective regularization method that is widely used in practice. Its implementation requires minimal changes to the model architecture and can be easily integrated into most neural networks. However, choosing the right probability value p is an important aspect of this method. If the probability is too large, it can lead to a significant decrease in the model's performance on normal data. On the other hand, too small a probability may not provide sufficient protection against overfitting and adversarial attacks. Dropout is an active method, as it modifies the model during training, making it more resilient to various types of attacks.

L2 Regularization is a regularization method based on adding a penalty for large values of model weights [148]. The idea of the method is shown in the following formula:

$$L(\theta) = \text{Loss}(\theta) + \lambda \sum_i \theta_i^2. \quad (21)$$

The formula describes a loss function that includes the main loss function $\text{Loss}(\theta)$ and a regularization term $\lambda \sum_i \theta_i^2$. The regularization term adds a penalty for large values of the weights θ, which helps control the complexity of the model and makes it more robust to adversarial attacks. This method is especially useful for models that are prone to overfitting and use too large weights to classify data.

The regularization coefficient λ plays a key role in this method. If λ is too large, it can lead to a significant decrease in the performance of the model on normal data. On the other hand, too small a value of λ may not provide sufficient protection against overfitting and adversarial attacks. L2 Regularization is an active method, as it modifies the model training process by adding a regularization term to the loss function.

Spectral Normalization is a method of normalizing the spectral radius of scales to control their scale [149]. The method is described by the following formula:

$$W' = \frac{W}{\sigma(W)}. \quad (22)$$

The formula describes the process of normalizing the W layer weights, where $\sigma(W)$ is the spectral radius, which is the maximum singular

value of the W matrix. This method normalizes the layer weights, which helps control their scale and improves the robustness of the model. Spectral Normalization is especially useful for models that are sensitive to the scale of the weights and may be vulnerable to adversarial attacks.

Calculating the spectral radius requires performing a singular value decomposition of the weight matrix, which can be computationally expensive. However, the results of this method can significantly improve the robustness of the model to various types of attacks. Spectral Normalization is an active method, as it modifies the model weights during training. This method requires a specialized approach and can be more difficult to implement compared to other methods.

Median Filtering is a method that replaces the value of each pixel with the median of the values in its neighborhood [150]. The method is described by the following formula:

$$I_{\text{filtered}} = \text{median}(I_{\text{window}}). \quad (23)$$

The formula describes the process of filtering an image. Here I_{window} is the filtering window of pixels around the current pixel, and $\text{median}(I_{\text{window}})$ is the median value of the pixels in the window.

The method helps remove small changes, such as noise or adversarial changes, making the model less susceptible to small changes. Median Filtering is a simple and effective method for removing small changes that can be used to generate adversarial examples. It does not require significant changes to the model architecture and can be easily integrated into an existing data processing pipeline. However, too large a filtering window can result in significant blurring of the image, which can also negatively affect the performance of the model. Median Filtering is an active method, as it modifies the input data to improve the robustness of the model.

5. Discussion. Summarizing the results of the conducted research, the following conclusions can be drawn.

1. Currently, there are a considerable number of papers that study adversarial attacks and protection methods against them. Moreover, the number of such articles increases over time. This suggests that the topic of protection against attacks on machine intelligence systems is highly relevant. Moreover, many papers are of a review nature. They systematize well-known and newly emerged attacks and protection methods, propose classification features and classification systems for attacks and protection methods. Therefore, numerous review articles on this topic, published between 2020 and 2025 in journals ranked in the Scopus Q1 quartile, were

selected as the initial data for the conducted research. A total of 30 such review articles were selected [19 – 48]. It was noted that, in general, the material devoted to the analysis of adversarial attacks and protection methods tends to be repeated across review papers from year to year. Therefore, such a seemingly small number of review papers analyzed in the conducted study seems to be quite sufficient.

2. In addition, papers in the field of adversarial attacks were selected and studied, which are not review-based but research-based and are devoted to the field of energy infrastructure. The analysis of these studies showed that, firstly, adversarial attacks can affect any elements of smart grids or energy systems in which ML models are used for monitoring and control. Secondly, new adversarial attacks on energy infrastructure are constantly appearing. Therefore, the search for defense methods remains crucial. Thirdly, most of the protection methods against adversarial attacks used in energy infrastructures are based on various ML models.

3. Analyzing the classification systems of adversarial attacks proposed in [19 – 48], it was noted that all of them were unique. They differed both in classification features and in the values of these features. However, in the course of the conducted study, we sought to identify the general trends inherent in the classification of adversarial attacks and to propose a generalized classification of such attacks that would cover other known classifications. As a result, in this paper, a set of eight most relevant classification features of adversarial attacks is proposed, which was tested on a set consisting of 31 of the most well-known attacks. The dominant classification feature is “knowledge of the attacked system”, which takes the values “white box”, “black box”, and “gray box”.

4. Analyzing the distribution of the selected attacks by the values of classification features, it was found that all attacks can be divided into groups with high, medium, and low popularity. To protect against attacks with high popularity, a universal approach can be used to detect them, since they are described by frequently used feature values. Attacks of medium popularity contain rarer features. Therefore, on the one hand, approaches from the previous group are partially suitable for their detection, and on the other hand, these attacks have their own specifics. Low-popularity attacks require separate individual approaches for their detection.

5. Analyzing the classification systems of protection methods against adversarial attacks proposed in [19 – 48], it was also noted that all of them are unique. Therefore, one of the goals of the study was to form a generalized classification of protection methods against such attacks. A set of six classification features was proposed, which was tested on a set consisting of 17 of the most well-known protection methods. As in the case

of the classification of attacks, an assessment of protection methods by their popularity was carried out, and it was proposed to distribute protection methods into two groups: high and low popularity. High-popularity methods have a wider scope of application and are effective against a wider range of adversarial attacks. At the same time, the implementation of these methods is usually simple. Low-popularity methods provide protection against a smaller number of attacks and are characterized by increased complexity.

6. The conducted study provided a brief description of adversarial attacks and protection methods against them, using, where possible, a formal description of these attacks and methods. Thus, a formal theoretical basis was laid for further research on assessing the capabilities of adversarial attacks and improving protection methods against them.

7. In conclusion, it is important to highlight some features inherent in the implementation of adversarial attacks and protection methods against them, identified in the course of the conducted study.

Thus, the One Pixel Attack becomes ineffective if two or more pooling layers are used in the basic design of the ML model. The specific type of pooling used is irrelevant, since the combination of layers of this type neutralizes the effect of pixel differences at the level of high-level features [151].

The JSMA attack has the following very important feature: it cannot function simultaneously with the ML model. Therefore, the correct organization of data flows in the pipeline for constructing an applied AI model will help to completely eliminate the negative effect of introducing JSMA as a malicious component [152]. In this case, no additional add-ons will be required to control the functioning of the main machine learning model.

It is also worth noting that defense against FA attacks is quite non-trivial: the ML model may malfunction when anomalous values appear in the data set [153]. Anomalous indicators include not only hacking attempts but also, most often, simply the rarest values. For example, an electrical voltage of 5,000 V may seem abnormal to the generalization ability of the algorithm if the model is most often operated with data of the average statistical voltage. For this reason, the implementation of defense components against FA attacks should be done with special caution, since non-standard data can be interpreted by the defense system as an attempt of unauthorized access to data.

Finally, it should be noted that the Defensive Distillation technique has its own characteristics when protecting ensemble models. Thus, if an algorithm with privileged information is used in the decision model, then

the separate functioning of the student model and the teacher model can lead to collisions during the normal operation of the base model.

It should be noted that the analysis of defense methods for ML systems in the energy domain allowed the authors to identify three most effective approaches: noise reduction, compression, and neural cleansing of input data. Preliminary experiments demonstrated the high efficiency of the approach based on a combination of these methods [154]. Optimization of the parameters of combining these methods allows us to almost completely eliminate the negative consequences of such popular attacks as FGSM, ZOO, and OPA. Further research in this direction is ongoing.

6. Conclusion. The article presents the results of a comprehensive analysis of attacks against ML systems and defense methods against them. Energy systems and networks, being an important element of the infrastructure, are increasingly dependent on digital technologies. Modern energy systems are no longer purely physical structures, such as power plants and transmission lines. They are becoming more complex and interconnected due to the introduction of digital technologies: automation, the Internet of Things, and big data. Their integration with artificial intelligence requires the introduction of more stringent security measures.

The presented classification of attack methods and techniques will allow developers to identify the most vulnerable components of such intelligent systems, as well as find a suitable defense method from those proposed in this article. The compilation of the material discussed in this article was carried out on the basis of the inclusion of the latest achievements in the field of information security of applied ML models. Therefore, the presented defense methods will be a relevant guide for developers of AI models. Separately, it is worth noting that the integration of applied AI models into applications may imply other risks to information security that were not described in this article, due to the potential vulnerabilities in the application's underlying software and hardware complex.

The study identified several trends in the direction of attacks against AI-based systems. The first is attacks using deep learning neural networks. With the increasing computing power and availability of equipment, there are more and more opportunities to carry out such attacks. They make models vulnerable to attacks with the distortion of input data. The second trend is subtle, difficult to distinguish attacks that involve making minimal changes to normal data, making their detection even more difficult. The third is compound attacks that occur in several stages. Attacks are becoming more sophisticated, using multi-layered strategies that may include preliminary training on real data before the attack.

Studying the vulnerabilities of AI systems and developing threat models is an important task for ensuring security in the context of rapid technological development and the connection of AI to energy systems, where reliable transmission and processing of real-time data plays a key role. The creation of active monitoring and anomaly detection systems can prevent malicious actions and reduce the risks associated with the use of AI technologies in energy infrastructure. It is important to focus on the need to improve knowledge in the field of artificial intelligence and cybersecurity, namely the training of IT specialists servicing energy facilities. In the context of the growing need for security measures, the development of methods and algorithms for identifying yet-unknown attacks using AI also remains important.

On November 7, 2024, Rafael Midkhatovich Yusupov, a prominent scientist in the fields of computer science, information technology, and control theory, founder and leader of scientific schools on the theoretical foundations of the informatization of society and on the sensitivity theory of complex information and control systems, Doctor of Engineering Sciences, Professor, Corresponding Member of the Russian Academy of Sciences, Honored Scientist and Engineer of the Russian Federation, Head of the Research Department at SPIIRAS, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Director of SPIIRAS (1991-2018), and Editor-in-Chief of the journal "Informatics and Automation," passed away at the age of 90. Rafael Yusupov actively promoted the field of cybersecurity and supported the work of the authors of this article.

References

1. Nascimento E.d.S., Ahmed I., Oliveira E., Palheta M.P., Steinmacher I., Conte T. Understanding Development Process of Machine Learning Systems: Challenges and Solutions. Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM). 2019. pp. 1–6.
2. Farhat R., Mourali Y., Jemni M., Ezzedine H. An overview of Machine Learning Technologies and their use in E-learning. Proceedings of the 2020 International Multi-Conference on Organization of Knowledge and Advanced Technologies (OCTA). 2020. pp. 1–4.
3. Liu Q., Li P., Zhao W., Cai W., Yu S., Leung V.C.M. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. IEEE Access. 2018. vol. 6. pp. 12103–12117.
4. Pitropakis N., Panaousis E., Giannetsos T., Anastasiadis E., Loukas G. A taxonomy and survey of attacks against machine learning. Computer Science Review. 2019. vol. 34.
5. Tcukanova O.A., Yarskaya A.A., Torosyan A.A. Artificial Intelligence as a New Stage in the Development of Business Intelligence Systems. Proceedings of the International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS). 2022. pp. 315–318.

6. Bland J.A., Petty M.D., Whitaker T.S., Maxwell K.P., Cantrell W.A. Machine Learning Cyberattack and Defense Strategies. Computers & Security. 2020. vol. 92.
7. Onyeji I., Bazilian M., Bronk C. Cyber security and critical energy infrastructure. The Electricity Journal. 2014. vol. 27. pp. 52–60.
8. Santis E.D., Rizzi A., Sadeghian A. A cluster-based dissimilarity learning approach for localized fault classification in smart grids. Swarm Evol. Comput. 2018. vol. 39. pp. 267–278.
9. Umapathy K., Dinesh Kumar T., Poojitha G., Khyathi Sri D., Pavaneeswar C., Amannah C. Machine Learning Applications for the Smart Grid. Data Analytics for Smart Grids Applications – A Key to Smart City Development. Intelligent Systems Reference Library. 2023. vol. 247. pp. 251–270.
10. Zhao Z., Chen G. An Overview of Cyber Security for Smart Grid. Proceedings of the 2018 IEEE 27th International Symposium on Industrial Electronics (ISIE). 2018. pp. 1127–1131.
11. Biggio B., Corona I., Maiorca D., Nelson B., Srndic N., Laskov P., Giacinto G., Roli F. Evasion Attacks against Machine Learning at Test Time. Advanced Information Systems Engineering. Lecture Notes in Computer Science. 2013. vol. 7908. pp. 387–402.
12. Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. Proceedings of the International Conference on Learning Representations (ICLR'15). 2015. arXiv preprint arXiv:1412.6572.
13. Nguyen T.N., Liu B.-H., Nguyen N.P., Chou J.-T. Cyber Security of Smart Grid: Attacks and Defenses. Proceedings of the IEEE International Conference on Communications (ICC). 2020. pp. 1–6.
14. Kawoosa A.I., Prashar D. A Review of Cyber Securities in Smart Grid Technology. Proceedings of the 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM). 2021. pp. 151–156.
15. Madhavarapu V.P.K., Bhattacharjee S., Dasy S.K. A Generative Model for Evasion Attacks in Smart Grid. Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 2022. pp. 1–6.
16. Cui L., Qu Y., Gao L., Xie G., Yu S. Detecting false data attacks using machine learning techniques in smart grid: A survey. J. Netw. Comput. Appl. 2020. vol. 170.
17. Kotenko I., Saenko I., Lauta O., Kribel K., Vasiliev N. Attacks Against Artificial Intelligence Systems: Classification, The Threat Model and the Approach to Protection. Proceedings of the Sixth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'22). 2023. pp. 293–302.
18. Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks Against Machine Learning Systems: Analysis and GAN-based Approach to Protection. Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). 2023. pp. 49–59.
19. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review. 2020. vol. 37.
20. Martins N., Cruz J.M., Cruz T., Henriques Abreu P. Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. IEEE Access. 2020. vol. 8. pp. 35403–35419.
21. Oseni A., Moustafa N., Janicke H., Liu P., Tari Z., Vasilakos A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. 2020. arXiv preprint arXiv:2102.04661.

22. Xu H., Ma Y., Liu H.C., Deb D., Liu H., Tang J.-L., Jain A.K. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Int. J. Autom. Comput. 2020. vol. 17. pp. 151–178.
23. Ren K., Zheng T., Qin Zh., Liu X. Adversarial Attacks and Defenses in Deep Learning. Engineering. 2020. vol. 6. pp. 346–360.
24. Zhou X., Canady R., Li Y., Koutsoukos X., Gokhale A. Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair. Dynamic Data Driven Applications Systems (DDDAS 2020). Lecture Notes in Computer Science. 2020. vol. 12312. pp. 102–109.
25. Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. 2021. vol. 9. pp. 155161–155196.
26. Zhang H., Liu B., Wu H. Smart Grid Cyber-Physical Attack and Defense: A Review. IEEE Access. 2021. vol. 9. pp. 29641–29659.
27. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol. 2021. vol. 6. pp. 25–45.
28. Rosenberg I., Shabtai A., Elovici Y., Rokach L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Comput. Surv. 2021. vol. 54(5). pp. 1–36.
29. Tian J., Wang B., Li J., Konstantinou C. Adversarial attack and defense methods for neural network-based state estimation in smart grid. IET Renewable Power Generation. 2022. vol. 16. pp. 3507–3518.
30. Kong Z., Xue J., Wang Y., Huang L., Niu Z., Li F., Meng W. A Survey on Adversarial Attack in the Age of Artificial Intelligence. Wireless Communications and Mobile Computing. 2021. vol. 2021(1).
31. Zhou Sh., Liu Ch., Ye D., Zhu T., Zhou W., Yu Ph.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. ACM Comput. Surv. 2022. vol. 55(8). pp. 1–39.
32. Khamaiseh S.Y., Bagagem D., Al-Alaj A., Mancino M., Alomari H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. IEEE Access. 2022. vol. 10. pp. 102266–102291.
33. Liang H., He E., Zhao Y., Jia Z., Li H. Adversarial Attack and Defense: A Survey. Electronics. 2022. vol. 11.
34. Tian Q., Zhang S., Mao Sh., Lin Y. Adversarial attacks and defenses for digital communication signals identification. Digital Communications and Networks. 2024. vol. 10. no. 3. pp. 756–764.
35. Anastasiou Th., Karagiorgou S., Petrou P., Papamartzivanos D., Giannetsos Th., Tsirigotaki G., Keizer J. Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems. Sensors. 2022. vol. 22.
36. Li Y., Cheng M., Hsieh Ch.-J., Lee Th.C.M. A Review of Adversarial Attack and Defense for Classification Methods. The American Statistician. 2022. vol. 76(4). pp. 329–345.
37. Tian J., Wang B., Li J., Wang Z. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid. IEEE Transactions on Network Science and Engineering. 2021. vol. 9(2). pp. 807–819.
38. Li H., Namiot D. A Survey of Adversarial Attacks and Defenses for Image Data on Deep Learning. International Journal of Open Information Technologies. 2022. vol. 10. pp. 9–16.
39. Girdhar M., Hong J., Moore J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology. 2023. vol. 4. pp. 417–437.

40. Goyal Sh., Doddapaneni S., Khapra M.M., Ravindran B. A Survey of Adversarial Defenses and Robustness in NLP. ACM Comput. Surv. 2023. vol. 55.
41. Al-Khassawneh Y.A. A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges. Indonesian Journal of Science and Technology. 2023. vol. 8. pp. 79–96.
42. He K., Kim D.D., Asghar M.R. Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. IEEE Communications Surveys & Tutorials. 2023. vol. 25(1). pp. 538–566.
43. Sun L., Dou Y., Yang C., Zhang K., Wang J., Yu Ph.S., He L., Li B. Adversarial Attack and Defense on Graph Data: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35(8). pp. 7693–7711.
44. Qureshi A.U.H., Larijani H., Yousefi M., Adeel A., Mtetwa N. An Adversarial Approach for Intrusion Detection Systems Using Jacobian Saliency Map Attacks (JSMA) Algorithm. Computers. 2020. vol. 9(3).
45. Hoang V.-T., Ergu Y.A., Nguyen V.-L., Chang R.-G. Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey. Journal of Network and Computer Applications. 2024. vol. 232.
46. Baniecki H., Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion. 2024. vol. 107.
47. Andrade E. de O., Guérin J., Viterbo J., Sampaio I.G.B. Adversarial attacks and defenses in person search: A systematic mapping study and taxonomy. Image and Vision Computing. 2024. vol. 148.
48. Zhang Ch., Hu M., Li W., Wang L. Adversarial attacks and defenses on text-to-image diffusion models: A survey. Information Fusion. 2025. vol. 114.
49. Hao J., Tao Y. Adversarial attacks on deep learning models in smart grids. Energy Reports. 2022. vol. 8. pp. 123–129.
50. Sande-Ríos J., Canal-Sánchez J., Manzano-Hernández C., Pastrana S. Threat Analysis and Adversarial Model for Smart Grids. Proceeding of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). 2024. pp. 130–145.
51. Elsisi M., Su C.-L., Ali M.N. Design of Reliable IoT Systems With Deep Learning to Support Resilient Demand Side Management in Smart Grids Against Adversarial Attacks. Transactions on Industry Applications. 2024. vol. 60. pp. 2095–2106.
52. Ahmadian S., Malki H., Han Z. Cyber Attacks on Smart Energy Grids Using Generative Adverserial Networks. Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP). 2018. pp. 942–946.
53. Takiddin A., Ismail M., Zafar U., Serpedin E. Robust Electricity Theft Detection Against Data Poisoning Attacks in Smart Grids. IEEE Transactions on Smart Grid. 2021. vol. 12. pp. 2675–2684.
54. Takiddin A., Ismail M., Serpedin E. Robust Data-Driven Detection of Electricity Theft Adversarial Evasion Attacks in Smart Grids. IEEE Transactions on Smart Grid. 2023. vol. 14. pp. 663–676.
55. Bondok A.H., Mahmoud M., Badr M.M., Fouda M.M., Abdallah M., Alsabaan M. Novel Evasion Attacks against Adversarial Training Defense for Smart Grid Federated Learning. IEEE Access. 2023. vol. 11. pp. 112953–112972.
56. Sampedro G.A., Ojo S., Krichen M., Alamro M.A., Mihoub A., Karovic V. Defending AI Models Against Adversarial Attacks in Smart Grids Using Deep Learning. IEEE Access. 2024. vol. 12. pp. 157408–157417. DOI: 10.1109/ACCESS.2024.3473531.
57. Khan S.U., Mynuddin M., Nabil M. AdaptEdge: Targeted Universal Adversarial Attacks on Time Series Data in Smart Grids. IEEE Transactions on Smart Grid. 2024. vol. 15. pp. 5072–5086.

58. Berghout T., Benbouzid M., Amirat Y. Towards Resilient and Secure Smart Grids against PMU Adversarial Attacks: A Deep Learning-Based Robust Data Engineering Approach. Electronics. 2023. vol. 12.
59. Teryak H., Albaseer A., Abdallah M., Al-Kuwari S., Qaraqe M. Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids. IEEE Open Journal of the Industrial Electronics Society. 2023. vol. 4. pp. 629–642.
60. Ness S. Adversarial Attack Detection in Smart Grids Using Deep Learning Architectures. IEEE Access. 2025. vol. 13. pp. 16314–16323. DOI: 10.1109/ACCESS.2024.3523409.
61. Zhang G., Sikdar B. Ensemble and Transfer Adversarial Attack on Smart Grid Demand-Response Mechanisms. Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). IEEE, 2022. pp. 53–58.
62. Zhang G., Sikdar B. A Novel Adversarial FDI Attack and Defense Mechanism for Smart Grid Demand-Response Mechanisms. IEEE Transactions on Industrial Cyber-Physical Systems. 2024. vol. 2. pp. 380–390. DOI: 10.1109/TICPS.2024.3448380.
63. Reda H.T., Anwar A., Mahmood A. Comprehensive survey and taxonomies of false data injection attacks in smart grids: attack models, targets, and impacts. Renewable and Sustainable Energy Reviews. 2022. vol. 163.
64. Ren Y., Zhang H., Yang W., Li M., Zhang J., Li H. Transferable Adversarial Attack Against Deep Reinforcement Learning-Based Smart Grid Dynamic Pricing System. IEEE Transactions on Industrial Informatics. 2024. vol. 20. pp. 9015–9025.
65. He J., Xiang T., Wu T., Chen Z., Wang N., Guo S. Maintaining Privacy in Smart Grid: Utilizing the Adversarial Attack Paradigm to Counter Non-Intrusive Load Monitoring Models. IEEE Internet of Things Journal. 2024.
66. Guihai Z., Sikdar B. Adversarial Machine Learning Against False Data Injection Attack Detection for Smart Grid Demand Response. Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2021. pp. 352–357.
67. Aurangzeb M., Wang Y., Iqbal S., Naveed A., Ahmed Z., Alenezi M., Shouran M. Enhancing cybersecurity in smart grids: Deep black box adversarial attacks and quantum voting ensemble models for blockchain privacy-preserving storage. Energy Reports. 2024. vol. 11. pp. 2493–2515.
68. Melendez K.A., Matamala Y. Adversarial attacks in demand-side electricity markets. Applied Energy. 2025. vol. 377.
69. Nguyen T., Wang S., Alhazmi M., Nazemi M., Estebsari A., Dehghanian P. Electric Power Grid Resilience to Cyber Adversaries: State of the Art. IEEE Access. 2020. vol. 8. pp. 87592–87608.
70. Efatinasab E., Brighente A., Rampazzo M., Azadi N., Conti M. GAN-GRID: A Novel Generative Attack on Smart Grid Stability Prediction. Computer Security – ESORICS 2024. Lecture Notes in Computer Science. 2024. vol. 14982. pp. 374–393.
71. Takiddin A., Ismail M., Atat R., Serpedin E. Spatio-temporal Graph-Based Generation and Detection of Adversarial False Data Injection Evasion Attacks in Smart Grids. IEEE Transactions on Artificial Intelligence. 2024. vol. 5(12). pp. 6601–6616. DOI: 10.1109/TAI.2024.3464511.
72. Omara A., Kantarci B. An AI-driven solution to prevent adversarial attacks on mobile Vehicle-to-Microgrid services. Simulation Modelling Practice and Theory. 2024. vol. 137.
73. Tian J., Shen C., Wang B., Ren C., Xia X., Dong R., Cheng T. EVADE: Targeted Adversarial False Data Injection Attacks for State Estimation in Smart Grid. IEEE Transactions on Sustainable Computing. 2024.

74. Efatinasab E., Sinigaglia A., Azadi N., Susto G.A., Rampazzo M. Adversarially Robust Fault Zone Prediction in Smart Grids With Bayesian Neural Networks. IEEE Access. 2024. vol. 12. pp. 121169–121184.
75. Zhang Z., Liu M., Sun M., Deng R., Cheng P., Niyato D., Chow M.-Y., Chen J. Vulnerability of Machine Learning Approaches Applied in IoT-Based Smart Grid: A Review. IEEE Internet of Things Journal. 2024. vol. 11(11). pp. 18951–18975.
76. Sánchez G., Elbez G., Hagenmeyer V. Attacking Learning-based Models in Smart Grids: Current Challenges and New Frontiers. Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (e-Energy '24). 2024. pp. 589–595.
77. Huang R., Li Y. Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System. IEEE Transactions on Smart Grid. 2022. vol. 14(3). pp. 2367–2376.
78. Ruan J., Wang Q., Chen S., Lyu H., Liang G., Zhao J., Dong Z.Y. On Vulnerability of Renewable Energy Forecasting: Adversarial Learning Attacks. IEEE Transactions on Industrial Informatics. 2023. vol. 20(3). pp. 3650–3663.
79. Heinrich R., Scholz C., Vogt S., Lehna M. Targeted adversarial attacks on wind power forecasts. Mach Learn. 2024. vol. 113(2). pp. 863–889.
80. Tang N., Mao S., Nelms R.M. Adversarial Attacks to Solar Power Forecast. Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM). 2021. pp. 1–6.
81. KhanS.U., Mynuddin M., Adom I., Mahmoud M.N. Mitigating Targeted Universal Adversarial Attacks on Time Series Power Quality Disturbances Models. Proceedings of the 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2023. pp. 91–100.
82. Gunn S., Jang D., Paradise O., Spangher L., Spanos C.J. Adversarial poisoning attacks on reinforcement learning-driven energy pricing. Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22). 2024. pp. 262–265.
83. Zhang L., Jiang C., Pang A., He Y. Super-efficient detector and defense method for adversarial attacks in power quality classification. Applied Energy. 2024. vol. 361.
84. Sayghe A., Zhao J., Konstantinou C. Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation. Proceedings of the IEEE Power & Energy Society General Meeting (PESGM). 2020. pp. 1–5.
85. Akter K., Rahman M.A., Islam R.M., Sheikh R.I., Hossain M.J. Attack-resilient framework for wind power forecasting against civil and adversarial attacks. Electric Power Systems Research. 2025. vol. 238.
86. Li J., Wang J., Chen L., Yu Y. Defending Against Adversarial Attacks by Energy Storage Facility. Proceedings of the IEEE Power & Energy Society General Meeting (PESGM). 2022. pp. 1–5.
87. Farajzadeh-Zanjani M., Hallaji E., Razavi-Far R., Saif M., Parvania M. Adversarial Semi-Supervised Learning for Diagnosing Faults and Attacks in Power Grids. IEEE Transactions on Smart Grid. 2021. vol. 12(4). pp. 3468–3478.
88. Zhou Y., Ding Z., Wen Q., Wang Y. Robust Load Forecasting Towards Adversarial Attacks via Bayesian Learning. IEEE Transactions on Power Systems. 2023. vol. 38. pp. 1445–1459.
89. Zhang L., Jiang C., Chai Z., He Y. Adversarial attack and training for deep neural network based power quality disturbance classification. Engineering Applications of Artificial Intelligence. 2024. vol. 127.
90. Li W., Deka D., Wang R., Paternina M.R.A. Physics-Constrained Adversarial Training for Neural Networks in Stochastic Power Grids. IEEE Transactions on

Artificial Intelligence. 2024. vol. 5(3). pp. 1121–1131. DOI: 10.1109/TAI.2023.3236377.
91. Afrin A., Ardakanian O. Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems. Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy '23). 2023. pp. 446–458.
92. Tian J., Wang B., Wang Z., Cao K., Li J., Ozay M. Joint Adversarial Example and False Data Injection Attacks for State Estimation in Power Systems. IEEE Transactions on Cybernetics. 2022. vol. 52(12). pp. 13699–13713.
93. Ali M.N., Amer M., Elsisi M. Reliable IoT Paradigm with Ensemble Machine Learning for Faults Diagnosis of Power Transformers Considering Adversarial Attacks. IEEE Transactions on Instrumentation and Measurement. 2023. vol. 72. pp. 1–13.
94. Kim B., Shi Y., Sagduyu Y.E., Erpek T., Ulukus S. Adversarial Attacks against Deep Learning Based Power Control in Wireless Communications. Proceedings of the IEEE Globecom Workshops (GC Wkshps). 2021. pp. 1–6.
95. Manoj B.R., Sadeghi M., Larsson E.G. Adversarial Attacks on Deep Learning Based Power Allocation in a Massive MIMO Network. Proceedings of the ICC – IEEE International Conference on Communications. 2021. pp. 1–6.
96. Manoj B.R., Sadeghi M., Larsson E.G. Downlink Power Allocation in Massive MIMO via Deep Learning: Adversarial Attacks and Training. IEEE Transactions on Cognitive Communications and Networking. 2022. vol. 8(2). pp. 707–719.
97. Santos P.M., Manoj B.R., Sadeghi M., Larsson E.G. Universal Adversarial Attacks on Neural Networks for Power Allocation in a Massive MIMO System. IEEE Wireless Communications Letters. 2022. vol. 11(1). pp. 67–71.
98. Shabbir A., Manzoor H.U., Ahmed R.A., Halim Z. Resilience of Federated Learning Against False Data Injection Attacks in Energy Forecasting. Proceedings of the 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST). 2024. pp. 245–249.
99. Xu S., Yu L., Lin X. Robust Low-Overhead Control of DER Reactive Power under Adversarial Attacks and Uncertainty. Proceedings of the ICC 2024 – IEEE International Conference on Communications. 2024. pp. 3097–3103.
100. Kuzlu M., Sarp S., Catak F.O., Cali U., Zhao Y., Elma O., Guler O. Analysis of deceptive data attacks with adversarial machine learning for solar photovoltaic power generation forecasting. Electr. Eng. 2024. vol. 106(2). pp. 1815–1823.
101. Zideh M.J., Khalghani M.R., Solanki S.K. An unsupervised adversarial autoencoder for cyber attack detection in power distribution grids. Electric Power Systems Research. 2024. vol. 232.
102. Zeng L., Qiu D., Sun M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. Applied Energy. 2022. vol. 324.
103. Zografopoulos I., Hatziargyriou N.D., Konstantinou C. Distributed Energy Resources Cybersecurity Outlook: Vulnerabilities, Attacks, Impacts, and Mitigations. IEEE Systems Journal. 2023. vol. 17. pp. 6695–6709.
104. Santana E.J., Silva R.P., Zarpelão B.B., Barbon Junior S. Detecting and Mitigating Adversarial Examples in Regression Tasks: A Photovoltaic Power Generation Forecasting Case Study. Information. 2021. vol. 12(10).
105. Youssef E.-N.S., Labeau F., Kassouf M. Adversarial Dynamic Load-Altering Cyberattacks Against Peak Shaving Using Residential Electric Water Heaters. IEEE Transactions on Smart Grid. 2024. vol. 15(2). pp. 2073–2088.
106. Bhattacharjee A., Bai G., Tushar W., Verma A., Mishra S., Saha T.K. DeeBBAA: A Benchmark Deep Black-Box Adversarial Attack Against Cyber–Physical Power

Systems. IEEE Internet of Things Journal. 2024. vol. 11(24). pp. 40670–40688. DOI: 10.1109/JIOT.2024.3454257.
107. Genç D., Özuysal M., Tomur E. A Taxonomic Survey of Model Extraction Attacks. Proceedings of the 2023 IEEE International Conference on Cyber Security and Resilience (CSR). 2023. pp. 200–205.
108. Fan J., Yan Q., Li M., Qu G., Xiao Y. A Survey on Data Poisoning Attacks and Defenses. Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). 2022. pp. 48–55.
109. Wang S., Ko R.K.L., Bai G., Dong N., Choi T., Zhang Y. Evasion Attack and Defense on Machine Learning Models in Cyber-Physical Systems: A Survey. IEEE Communications Surveys & Tutorials. 2024. vol. 26. pp. 930–966.
110. El-Toukhy A.T., Mahmoud M., Bondok A.H., Fouda M.M., Alsabaan M. Evasion Attacks in Smart Power Grids: A Deep Reinforcement Learning Approach. Proceedings of the 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC). 2024. pp. 708–713.
111. Ali Alatwi H., Morisset C. Threat Modeling for Machine Learning-Based Network Intrusion Detection Systems. Proceedings of the IEEE International Conference on Big Data (Big Data). 2022. pp. 4226–4235.
112. Qiu S., Liu Q., Zhou S., Wu C. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied sciences. 2019. vol. 9(5).
113. Tuna O.F., Catak F.O., Eskil M.T. Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples. Multimed Tools Appl. 2022. vol. 81. pp. 11479–11500.
114. Dubrovkin J. Evaluation of undetectable perturbations of the peak parameters estimated by the least square curve fitting of analytical signal consisting of overlapping peaks. Chemom. and Intel. Lab. Syst. 2016. vol. 153. pp. 9–21.
115. Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 2002. vol. 86(11). pp. 2278–2324.
116. Liu X., Cheng M., Zhang H., Hsieh C.J. Towards Robust Neural Networks via Random Self-ensemble. Computer Vision – ECCV 2018. Lecture Notes in Computer Science. 2018. vol. 11211. pp. 381–397.
117. McCarthy A., Ghadafi E., Andriotis P., Legg P. Functionality-Preserving Adversarial Machine Learning for Robust Classification in Cybersecurity and Intrusion Detection Domains: A Survey. Journal of Cybersecurity and Privacy. 2022. vol. 2(1). pp. 154–190.
118. Goodfellow I., McDaniel P., Papernot N. Making Machine Learning Robust Against Adversarial Inputs. Communications of the ACM. 2018. vol. 61(7). pp. 56–66.
119. Li J., Yang Y., Sun J., Tomsovic K.L., Qi H. Towards Adversarial-Resilient Deep Neural Networks for False Data Injection Attack Detection in Power Grids. Proceedings of the 32nd International Conference on Computer Communications and Networks (ICCCN). 2023. pp. 1–10.
120. Li Y., Wang Y. Defense Against Adversarial Attacks in Deep Learning. Applied Sciences. 2018. vol. 9(1).
121. Zantedeschi V., Nicolae M.-I., Rawat A. Efficient Defenses Against Adversarial Attacks. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17). 2017. pp. 39–49.
122. Sarker A., Shen H., Sen T., Mendelson Q. Efficient Black-Box Adversarial Attacks for Deep Driving Maneuver Classification Models. Proceedings of the 18th International Conference on Mobile Ad Hoc and Smart Systems. 2021. pp. 536–544.
123. Gibert D., Zizzo G., Le Q., Planes J. Adversarial Robustness of Deep Learning-Based Malware Detectors via (De)Randomized Smoothing, IEEE Access. 2024. vol. 12. pp. 61152–61162.

124. Yuan X., He P., Zhu Q., Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. IEEE Transactions on Neural Networks and Learning Systems. 2019. vol. 30(9). pp. 2805–2824.
125. Ozturk E., Mesut A. Comparison of Learned Image Compression Methods and JPEG. Proceedings of the 2024 Innovations in Intelligent Systems and Applications Conference (ASYU). 2024. pp. 1–6.
126. Zhou Z., Firestone C. Humans can decipher adversarial images. Nat. Commun. 2019. vol. 10(1).
127. Mohammed A., Ali Z., Ahmad I. Enhancing adversarial robustness with randomized interlayer processing. Expert Systems with Applications. 2024. vol. 245.
128. Shi Y., Zeng H., Nguyen T.T. Adversarial Machine Learning for Network Security. Proceedings of the International Symposium on Technologies for Homeland Security. 2019. pp. 1–7.
129. Eleftheriadis C., Symeonidis A., Katsaros P. Adversarial robustness improvement for deep neural networks. Machine Vision and Applications. 2024. vol. 35(3).
130. Han S., Lin Ch., Shen Ch., Wang Q., Guan X. Interpreting Adversarial Examples in Deep Learning: A Review. ACM Comput. Surv. 2023. vol. 55. no. 14s. pp. 1–38.
131. Li B., Liu W. WAT: Improve the Worst-Class Robustness in Adversarial Training. In Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37(12). pp. 14982–14990.
132. Liu N., Du M., Guo R., Liu H., Hu X. Adversarial Attacks and Defenses: An Interpretation Perspective. SIGKDD Explor. Newsl. 2021. vol. 23(1). pp. 86–99.
133. Peng Y., Fu G., Luo Y., Yu Q., Wang L. CNN-based Steganalysis Detects Adversarial Steganography via Adversarial Training and Feature Squeezing. Proceedings of the 4th International Conference on Information Science, Parallel and Distributed Systems (ISPDS). 2023. pp. 165–169.
134. Wang W., Park Y., Lee T., Molloy I., Tang P., Xiong L. Utilizing Multimodal Feature Consistency to Detect Adversarial Examples on Clinical Summaries. Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics. 2020. pp. 259–268.
135. Shao J., Geng S., Fu Z., Xu W., Liu T., Hong S. CardioDefense: Defending against adversarial attack in ECG classification with adversarial distillation training. Biomedical Signal Processing and Control. 2024. vol. 91. DOI: 10.1016/j.bspc.2023.105922.
136. Li H., Xu X., Zhang X., Yang Sh., Li B. Qeba: Query-efficient boundary-based blackbox attack. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 1221–1230.
137. RyouD Ha I., Yoo H., Kim D., Han B. Robust Image Denoising Through Adversarial Frequency Mixup. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. pp. 2723–2732.
138. Tyukin I.Y., Higham D.J., Gorban A.N. On Adversarial Examples and Stealth Attacks in Artificial Intelligence Systems. Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2020. pp. 1–6.
139. Shimizu Y. Automatic Design System with Generative Adversarial Network and Vision Transformer for Efficiency Optimization of Interior Permanent Magnet Synchronous Motor. IEEE Transactions on Industrial Electronics. 2024. vol. 71(11). pp. 14600–14609.
140. Barth A., Rubinstein B.I.P., Sundararajan M., Mitchell J.C., Song D., Bartlett P.L. A Learning-Based Approach to Reactive Security. IEEE Transactions on Dependable and Secure Computing. 2012. vol. 9(4). pp. 482–493.

141. Aly A., Iqbal S., Youssef A., Mansour E. MEGR-APT: A Memory-Efficient APT Hunting System Based on Attack Representation Learning. IEEE Transactions on Information Forensics and Security. 2024. vol. 19. pp. 5257–5271.
142. Deng Y., Zheng X., Zhang T., Chen C., Lou G., Kim M. An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom). 2020. pp. 1–10.
143. He X., Yao J., Wang Y., Tang Z., Cheung K.C., See S., Han B., Chu X. NAS-LID: Efficient Neural Architecture Search with Local Intrinsic Dimension. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37(6). pp. 7839–7847.
144. Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18). 2018. pp. 2154–2156.
145. Kwatra S., Torra V. Data Reconstruction Attack against Principal Component Analysis. Security and Privacy in Social Networks and Big Data (SocialSec 2023). Lecture Notes in Computer Sciences. 2023. vol. 14097. pp. 79–92.
146. Pachika S., Reddy A.B., Pachika B., Karnam A. Generative Adversarial Networks: Overview. Proceedings of the Fifth International Conference on Computer and Communication Technologies (IC3T 2023). Lecture Notes in Computer Sciences. 2024. vol. 897. pp. 319–328.
147. Li H., Yu W., Huang H. Strengthening transferability of adversarial examples by adaptive inertia and amplitude spectrum dropout. Neural Networks. 2023. vol. 165. pp. 925–937.
148. Chen Z. Robust Sparse Online Learning through Adversarial Sparsity Constraints. Proceedings of the 2024 9th IEEE International Conference on Smart Cloud (SmartCloud). 2024. pp. 42–47.
149. Liu X., Chen X., Cheng J., Zhou L., Chen L., Li C., Zu S. Simulation of Complex Geological Architectures Based on Multistage Generative Adversarial Networks Integrating with Attention Mechanism and Spectral Normalization. IEEE Transactions on Geoscience and Remote Sensing. 2023. vol. 61. pp. 1–15.
150. Ahmed S., Islam S. Methods in detection of median filtering in digital images: a survey. Multimed. Tools Appl. 2023. vol. 82. pp. 43945–43965.
151. Elderman R., Pater L.J.J., Thie A.S., Drugan M.M., Wiering M.A. Adversarial Reinforcement Learning in a Cyber Security Simulation. Proceedings of the 9th International Conference on Agents and Artificial Intelligence. 2017. pp. 559–566.
152. Ling Y., Yong Z., Pengfei W. Euclidean and Rapid Jacobian-based Saliency Maps Attacks. Proceedings of the 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). 2021. pp. 355–361.
153. Vorobeychik Y., Kantarcioglu M. Introduction. Adversarial Machine Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2018. pp. 1–4.
154. Kotenko I., Saenko I., Lauta O., Vasiliev N., Sadovnikov V. A Noise-Based Approach Augmented with Neural Cleanse and JPEG Compression to Counter Adversarial Attacks against Image Classification Systems. Proceedings of the 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP). 2025. pp. 576–583.

Kotenko Igor — Ph.D., Dr.Sci., Professor, Honored scientist of the russian federation, Chief researcher, head of the laboratory, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: computer network security, including security policy management, access control, authentication, security analysis, detection of computer attacks, firewalls, protection against viruses and network worms, analysis and verification of security protocols and information

security systems, protection of software from hacking and digital rights management, modeling and visualization technologies to counter cyberterrorism. The number of publications — 1000. ivkote@comsec.spb.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-7181; fax: +7(812)328-4450.

Saenko Igor — Ph.D., Dr.Sci., Professor, Chief researcher, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: automated information systems, information security, data processing and transmission via communication channels, modeling theory and mathematical statistics, information theory. The number of publications — 500. ibsaen@comsec.spb.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328–7181; fax: +7(812)328-4450.

Lauta Oleg — Ph.D., Dr.Sci., Associate Professor, Professor of the department, Department of comprehensive information security, Admiral Makarov State University of Maritime and Inland Shipping. Research interests: protection against computer attacks. The number of publications — 200. laos-82@yandex.ru; 5/7, Dvinskaya St., 198035, St. Petersburg, Russia; office phone: +7(911)842-0228.

Sadovnikov Vladimir — Postgraduate student, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: protection against computer attacks. The number of publications — 25. bladimir1998@mail.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-7181; fax: +7(812)328-4450.

Ichetovkin Egor — postgraduate student, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: protection against computer attacks. The number of publications — 7. egor.email@list.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-7181; fax: +7(812)328-4450.

Li Wei — Professor, College of computer science and technology, Harbin Engineering University. Research interests: protection against computer attacks, machine learning, artificial intelligence. The number of publications — 200. wei.li@hrbeu.edu.cn; 145, улица Наньтун, район Нанган, 150001, Harbin, China; office phone: 0086(0451)8251921.

Acknowledgements. The research is partially supported by the budget theme FFZF-2025-0016.

УДК 004.056.5 DOI 10.15622/ia.24.6.8

И.В. Котенко, И.Б. Саенко, О.С. Лаута, В.Е. Садовников,
Е.А. Ичетовкин, В. Ли

АНАЛИЗ СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ ПО ЗАЩИТЕ ОТ СОСТЯЗАТЕЛЬНЫХ АТАК В ЭНЕРГЕТИЧЕСКИХ СИСТЕМАХ

Котенко И.В., Саенко И.Б., Лаута О.С., Садовников В.Е., Ичетовкин Е.А., Ли В. **Анализ современных исследований по защите от состязательных атак в энергетических системах.**

Аннотация. Системы на основе машинного обучения в настоящее время являются привлекательными мишенями для злоумышленников, поскольку нарушение работы таких систем может иметь серьезные последствия для объектов критической инфраструктуры, в частности, энергетических систем. В связи с этим количество различных типов кибератак на системы машинного обучения, которые называются состязательными атаками, постоянно растёт, и эти атаки являются предметом изучения многих исследователей. Соответственно, ежегодно появляется множество публикаций, посвящённых обзорам состязательных атак и методов защиты от них. Многие виды состязательных атак и методы защиты в этих обзорных статьях пересекаются. Однако в более поздних исследованиях содержится информация о новых типах атак и методах защиты. Цель данной статьи – проанализировать исследования, проведённые за последние шесть лет и опубликованные в высокорейтинговых журналах, с акцентом на обзорные работы. Результатом исследования является уточнённая классификация состязательных атак, характеристика наиболее распространённых атак, а также уточнённая классификация и характеристика методов защиты от этих атак. Основное внимание в анализе уделяется состязательным атакам, нацеленным на энергетические системы. В заключительной части статьи рассматриваются преимущества и недостатки различных методов противодействия состязательным атакам.

Ключевые слова: кибератаки, искусственный интеллект, машинное обучение, состязательные атаки, модель угроз, методы защиты, обзор, энергетические системы, классификация.

Литература

1. Nascimento E.d.S., Ahmed I., Oliveira E., Palheta M.P., Steinmacher I., Conte T. Understanding Development Process of Machine Learning Systems: Challenges and Solutions. Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM). 2019. pp. 1–6.
2. Farhat R., Mourali Y., Jemni M., Ezzedine H. An overview of Machine Learning Technologies and their use in E-learning. Proceedings of the 2020 International Multi-Conference on Organization of Knowledge and Advanced Technologies (OCTA). 2020. pp. 1–4.
3. Liu Q., Li P., Zhao W., Cai W., Yu S., Leung V.C.M. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. IEEE Access. 2018. vol. 6. pp. 12103–12117.
4. Pitropakis N., Panaousis E., Giannetsos T., Anastasiadis E., Loukas G. A taxonomy and survey of attacks against machine learning. Computer Science Review. 2019. vol. 34.
5. Tcukanova O.A., Yarskaya A.A., Torosyan A.A. Artificial Intelligence as a New Stage in the Development of Business Intelligence Systems. Proceedings of the

International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS). 2022. pp. 315–318.
6. Bland J.A., Petty M.D., Whitaker T.S., Maxwell K.P., Cantrell W.A. Machine Learning Cyberattack and Defense Strategies. Computers & Security. 2020. vol. 92.
7. Onyeji I., Bazilian M., Bronk C. Cyber security and critical energy infrastructure. The Electricity Journal. 2014. vol. 27. pp. 52–60.
8. Santis E.D., Rizzi A., Sadeghian A. A cluster-based dissimilarity learning approach for localized fault classification in smart grids. Swarm Evol. Comput. 2018. vol. 39. pp. 267–278.
9. Umapathy K., Dinesh Kumar T., Poojitha G., Khyathi Sri D., Pavaneeswar C., Amannah C. Machine Learning Applications for the Smart Grid. Data Analytics for Smart Grids Applications – A Key to Smart City Development. Intelligent Systems Reference Library. 2023. vol. 247. pp. 251–270.
10. Zhao Z., Chen G. An Overview of Cyber Security for Smart Grid. Proceedings of the 2018 IEEE 27th International Symposium on Industrial Electronics (ISIE). 2018. pp. 1127–1131.
11. Biggio B., Corona I., Maiorca D., Nelson B., Srndic N., Laskov P., Giacinto G., Roli F. Evasion Attacks against Machine Learning at Test Time. Advanced Information Systems Engineering. Lecture Notes in Computer Science. 2013. vol. 7908. pp. 387–402.
12. Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. Proceedings of the International Conference on Learning Representations (ICLR'15). 2015. arXiv preprint arXiv:1412.6572.
13. Nguyen T.N., Liu B.-H., Nguyen N.P., Chou J.-T. Cyber Security of Smart Grid: Attacks and Defenses. Proceedings of the IEEE International Conference on Communications (ICC). 2020. pp. 1–6.
14. Kawoosa A.I., Prashar D. A Review of Cyber Securities in Smart Grid Technology. Proceedings of the 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM). 2021. pp. 151–156.
15. Madhavarapu V.P.K., Bhattacharjee S., Dasy S.K. A Generative Model for Evasion Attacks in Smart Grid. Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 2022. pp. 1–6.
16. Cui L., Qu Y., Gao L., Xie G., Yu S. Detecting false data attacks using machine learning techniques in smart grid: A survey. J. Netw. Comput. Appl. 2020. vol. 170.
17. Kotenko I., Saenko I., Lauta O., Kribel K., Vasiliev N. Attacks Against Artificial Intelligence Systems: Classification, The Threat Model and the Approach to Protection. Proceedings of the Sixth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'22). 2023. pp. 293–302.
18. Kotenko I., Saenko I., Lauta O., Vasiliev N., Iatsenko D. Attacks Against Machine Learning Systems: Analysis and GAN-based Approach to Protection. Proceedings of the Seventh International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'23). 2023. pp. 49–59.
19. Huang X., Kroening D., Ruan W., Sharp J., Sun Y., Thamo E., Wu M., Yi X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review. 2020. vol. 37.
20. Martins N., Cruz J.M., Cruz T., Henriques Abreu P. Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. IEEE Access. 2020. vol. 8. pp. 35403–35419.
21. Oseni A., Moustafa N., Janicke H., Liu P., Tari Z., Vasilakos A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. 2020. arXiv preprint arXiv:2102.04661.

22. Xu H., Ma Y., Liu H.C., Deb D., Liu H., Tang J.-L., Jain A.K. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Int. J. Autom. Comput. 2020. vol. 17. pp. 151–178.
23. Ren K., Zheng T., Qin Zh., Liu X. Adversarial Attacks and Defenses in Deep Learning. Engineering. 2020. vol. 6. pp. 346–360.
24. Zhou X., Canady R., Li Y., Koutsoukos X., Gokhale A. Overcoming Stealthy Adversarial Attacks on Power Grid Load Predictions Through Dynamic Data Repair. Dynamic Data Driven Applications Systems (DDDAS 2020). Lecture Notes in Computer Science. 2020. vol. 12312. pp. 102–109.
25. Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. 2021. vol. 9. pp. 155161–155196.
26. Zhang H., Liu B., Wu H. Smart Grid Cyber-Physical Attack and Defense: A Review. IEEE Access. 2021. vol. 9. pp. 29641–29659.
27. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol. 2021. vol. 6. pp. 25–45.
28. Rosenberg I., Shabtai A., Elovici Y., Rokach L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Comput. Surv. 2021. vol. 54(5). pp. 1–36.
29. Tian J., Wang B., Li J., Konstantinou C. Adversarial attack and defense methods for neural network-based state estimation in smart grid. IET Renewable Power Generation. 2022. vol. 16. pp. 3507–3518.
30. Kong Z., Xue J., Wang Y., Huang L., Niu Z., Li F., Meng W. A Survey on Adversarial Attack in the Age of Artificial Intelligence. Wireless Communications and Mobile Computing. 2021. vol. 2021(1).
31. Zhou Sh., Liu Ch., Ye D., Zhu T., Zhou W., Yu Ph.S. Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity. ACM Comput. Surv. 2022. vol. 55(8). pp. 1–39.
32. Khamaiseh S.Y., Bagagem D., Al-Alaj A., Mancino M., Alomari H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. IEEE Access. 2022. vol. 10. pp. 102266–102291.
33. Liang H., He E., Zhao Y., Jia Z., Li H. Adversarial Attack and Defense: A Survey. Electronics. 2022. vol. 11.
34. Tian Q., Zhang S., Mao Sh., Lin Y. Adversarial attacks and defenses for digital communication signals identification. Digital Communications and Networks. 2024. vol. 10. no. 3. pp. 756–764.
35. Anastasiou Th., Karagiorgou S., Petrou P., Papamartzivanos D., Giannetsos Th., Tsirigotaki G., Keizer J. Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems. Sensors. 2022. vol. 22.
36. Li Y., Cheng M., Hsieh Ch.-J., Lee Th.C.M. A Review of Adversarial Attack and Defense for Classification Methods. The American Statistician. 2022. vol. 76(4). pp. 329–345.
37. Tian J., Wang B., Li J., Wang Z. Adversarial Attacks and Defense for CNN Based Power Quality Recognition in Smart Grid. IEEE Transactions on Network Science and Engineering. 2021. vol. 9(2). pp. 807–819.
38. Li H., Namiot D. A Survey of Adversarial Attacks and Defenses for Image Data on Deep Learning. International Journal of Open Information Technologies. 2022. vol. 10. pp. 9–16.
39. Girdhar M., Hong J., Moore J. Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models. IEEE Open Journal of Vehicular Technology. 2023. vol. 4. pp. 417–437.

40. Goyal Sh., Doddapaneni S., Khapra M.M., Ravindran B. A Survey of Adversarial Defenses and Robustness in NLP. ACM Comput. Surv. 2023. vol. 55.
41. Al-Khassawneh Y.A. A Review of Artificial Intelligence in Security and Privacy: Research Advances, Applications, Opportunities, and Challenges. Indonesian Journal of Science and Technology. 2023. vol. 8. pp. 79–96.
42. He K., Kim D.D., Asghar M.R. Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. IEEE Communications Surveys & Tutorials. 2023. vol. 25(1). pp. 538–566.
43. Sun L., Dou Y., Yang C., Zhang K., Wang J., Yu Ph.S., He L., Li B. Adversarial Attack and Defense on Graph Data: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35(8). pp. 7693–7711.
44. Qureshi A.U.H., Larijani H., Yousefi M., Adeel A., Mtetwa N. An Adversarial Approach for Intrusion Detection Systems Using Jacobian Saliency Map Attacks (JSMA) Algorithm. Computers. 2020. vol. 9(3).
45. Hoang V.-T., Ergu Y.A., Nguyen V.-L., Chang R.-G. Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey. Journal of Network and Computer Applications. 2024. vol. 232.
46. Baniecki H., Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion. 2024. vol. 107.
47. Andrade E. de O., Guérin J., Viterbo J., Sampaio I.G.B. Adversarial attacks and defenses in person search: A systematic mapping study and taxonomy. Image and Vision Computing. 2024. vol. 148.
48. Zhang Ch., Hu M., Li W., Wang L. Adversarial attacks and defenses on text-to-image diffusion models: A survey. Information Fusion. 2025. vol. 114.
49. Hao J., Tao Y. Adversarial attacks on deep learning models in smart grids. Energy Reports. 2022. vol. 8. pp. 123–129.
50. Sande-Ríos J., Canal-Sánchez J., Manzano-Hernández C., Pastrana S. Threat Analysis and Adversarial Model for Smart Grids. Proceeding of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). 2024. pp. 130–145.
51. Elsisi M., Su C.-L., Ali M.N. Design of Reliable IoT Systems With Deep Learning to Support Resilient Demand Side Management in Smart Grids Against Adversarial Attacks. Transactions on Industry Applications. 2024. vol. 60. pp. 2095–2106.
52. Ahmadian S., Malki H., Han Z. Cyber Attacks on Smart Energy Grids Using Generative Adverserial Networks. Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP). 2018. pp. 942–946.
53. Takiddin A., Ismail M., Zafar U., Serpedin E. Robust Electricity Theft Detection Against Data Poisoning Attacks in Smart Grids. IEEE Transactions on Smart Grid. 2021. vol. 12. pp. 2675–2684.
54. Takiddin A., Ismail M., Serpedin E. Robust Data-Driven Detection of Electricity Theft Adversarial Evasion Attacks in Smart Grids. IEEE Transactions on Smart Grid. 2023. vol. 14. pp. 663–676.
55. Bondok A.H., Mahmoud M., Badr M.M., Fouda M.M., Abdallah M., Alsabaan M. Novel Evasion Attacks against Adversarial Training Defense for Smart Grid Federated Learning. IEEE Access. 2023. vol. 11. pp. 112953–112972.
56. Sampedro G.A., Ojo S., Krichen M., Alamro M.A., Mihoub A., Karovic V. Defending AI Models Against Adversarial Attacks in Smart Grids Using Deep Learning. IEEE Access. 2024. vol. 12. pp. 157408–157417. DOI: 10.1109/ACCESS.2024.3473531.
57. Khan S.U., Mynuddin M., Nabil M. AdaptEdge: Targeted Universal Adversarial Attacks on Time Series Data in Smart Grids. IEEE Transactions on Smart Grid. 2024. vol. 15. pp. 5072–5086.

58. Berghout T., Benbouzid M., Amirat Y. Towards Resilient and Secure Smart Grids against PMU Adversarial Attacks: A Deep Learning-Based Robust Data Engineering Approach. Electronics. 2023. vol. 12.
59. Teryak H., Albaseer A., Abdallah M., Al-Kuwari S., Qaraqe M. Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids. IEEE Open Journal of the Industrial Electronics Society. 2023. vol. 4. pp. 629–642.
60. Ness S. Adversarial Attack Detection in Smart Grids Using Deep Learning Architectures. IEEE Access. 2025. vol. 13. pp. 16314–16323. DOI: 10.1109/ACCESS.2024.3523409.
61. Zhang G., Sikdar B. Ensemble and Transfer Adversarial Attack on Smart Grid Demand-Response Mechanisms. Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). IEEE, 2022. pp. 53–58.
62. Zhang G., Sikdar B. A Novel Adversarial FDI Attack and Defense Mechanism for Smart Grid Demand-Response Mechanisms. IEEE Transactions on Industrial Cyber-Physical Systems. 2024. vol. 2. pp. 380–390. DOI: 10.1109/TICPS.2024.3448380.
63. Reda H.T., Anwar A., Mahmood A. Comprehensive survey and taxonomies of false data injection attacks in smart grids: attack models, targets, and impacts. Renewable and Sustainable Energy Reviews. 2022. vol. 163.
64. Ren Y., Zhang H., Yang W., Li M., Zhang J., Li H. Transferable Adversarial Attack Against Deep Reinforcement Learning-Based Smart Grid Dynamic Pricing System. IEEE Transactions on Industrial Informatics. 2024. vol. 20. pp. 9015–9025.
65. He J., Xiang T., Wu T., Chen Z., Wang N., Guo S. Maintaining Privacy in Smart Grid: Utilizing the Adversarial Attack Paradigm to Counter Non-Intrusive Load Monitoring Models. IEEE Internet of Things Journal. 2024.
66. Guihai Z., Sikdar B. Adversarial Machine Learning Against False Data Injection Attack Detection for Smart Grid Demand Response. Proceedings of the IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2021. pp. 352–357.
67. Aurangzeb M., Wang Y., Iqbal S., Naveed A., Ahmed Z., Alenezi M., Shouran M. Enhancing cybersecurity in smart grids: Deep black box adversarial attacks and quantum voting ensemble models for blockchain privacy-preserving storage. Energy Reports. 2024. vol. 11. pp. 2493–2515.
68. Melendez K.A., Matamala Y. Adversarial attacks in demand-side electricity markets. Applied Energy. 2025. vol. 377.
69. Nguyen T., Wang S., Alhazmi M., Nazemi M., Estebsari A., Dehghanian P. Electric Power Grid Resilience to Cyber Adversaries: State of the Art. IEEE Access. 2020. vol. 8. pp. 87592–87608.
70. Efatinasab E., Brighente A., Rampazzo M., Azadi N., Conti M. GAN-GRID: A Novel Generative Attack on Smart Grid Stability Prediction. Computer Security – ESORICS 2024. Lecture Notes in Computer Science. 2024. vol. 14982. pp. 374–393.
71. Takiddin A., Ismail M., Atat R., Serpedin E. Spatio-temporal Graph-Based Generation and Detection of Adversarial False Data Injection Evasion Attacks in Smart Grids. IEEE Transactions on Artificial Intelligence. 2024. vol. 5(12). pp. 6601–6616. DOI: 10.1109/TAI.2024.3464511.
72. Omara A., Kantarci B. An AI-driven solution to prevent adversarial attacks on mobile Vehicle-to-Microgrid services. Simulation Modelling Practice and Theory. 2024. vol. 137.
73. Tian J., Shen C., Wang B., Ren C., Xia X., Dong R., Cheng T. EVADE: Targeted Adversarial False Data Injection Attacks for State Estimation in Smart Grid. IEEE Transactions on Sustainable Computing. 2024.

74. Efatinasab E., Sinigaglia A., Azadi N., Susto G.A., Rampazzo M. Adversarially Robust Fault Zone Prediction in Smart Grids With Bayesian Neural Networks. IEEE Access. 2024. vol. 12. pp. 121169–121184.
75. Zhang Z., Liu M., Sun M., Deng R., Cheng P., Niyato D., Chow M.-Y., Chen J. Vulnerability of Machine Learning Approaches Applied in IoT-Based Smart Grid: A Review. IEEE Internet of Things Journal. 2024. vol. 11(11). pp. 18951–18975.
76. Sánchez G., Elbez G., Hagenmeyer V. Attacking Learning-based Models in Smart Grids: Current Challenges and New Frontiers. Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (e-Energy '24). 2024. pp. 589–595.
77. Huang R., Li Y. Adversarial Attack Mitigation Strategy for Machine Learning-Based Network Attack Detection Model in Power System. IEEE Transactions on Smart Grid. 2022. vol. 14(3). pp. 2367–2376.
78. Ruan J., Wang Q., Chen S., Lyu H., Liang G., Zhao J., Dong Z.Y. On Vulnerability of Renewable Energy Forecasting: Adversarial Learning Attacks. IEEE Transactions on Industrial Informatics. 2023. vol. 20(3). pp. 3650–3663.
79. Heinrich R., Scholz C., Vogt S., Lehna M. Targeted adversarial attacks on wind power forecasts. Mach Learn. 2024. vol. 113(2). pp. 863–889.
80. Tang N., Mao S., Nelms R.M. Adversarial Attacks to Solar Power Forecast. Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM). 2021. pp. 1–6.
81. KhanS.U., Mynuddin M., Adom I., Mahmoud M.N. Mitigating Targeted Universal Adversarial Attacks on Time Series Power Quality Disturbances Models. Proceedings of the 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2023. pp. 91–100.
82. Gunn S., Jang D., Paradise O., Spangher L., Spanos C.J. Adversarial poisoning attacks on reinforcement learning-driven energy pricing. Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22). 2024. pp. 262–265.
83. Zhang L., Jiang C., Pang A., He Y. Super-efficient detector and defense method for adversarial attacks in power quality classification. Applied Energy. 2024. vol. 361.
84. Sayghe A., Zhao J., Konstantinou C. Evasion Attacks with Adversarial Deep Learning Against Power System State Estimation. Proceedings of the IEEE Power & Energy Society General Meeting (PESGM). 2020. pp. 1–5.
85. Akter K., Rahman M.A., Islam R.M., Sheikh R.I., Hossain M.J. Attack-resilient framework for wind power forecasting against civil and adversarial attacks. Electric Power Systems Research. 2025. vol. 238.
86. Li J., Wang J., Chen L., Yu Y. Defending Against Adversarial Attacks by Energy Storage Facility. Proceedings of the IEEE Power & Energy Society General Meeting (PESGM). 2022. pp. 1–5.
87. Farajzadeh-Zanjani M., Hallaji E., Razavi-Far R., Saif M., Parvania M. Adversarial Semi-Supervised Learning for Diagnosing Faults and Attacks in Power Grids. IEEE Transactions on Smart Grid. 2021. vol. 12(4). pp. 3468–3478.
88. Zhou Y., Ding Z., Wen Q., Wang Y. Robust Load Forecasting Towards Adversarial Attacks via Bayesian Learning. IEEE Transactions on Power Systems. 2023. vol. 38. pp. 1445–1459.
89. Zhang L., Jiang C., Chai Z., He Y. Adversarial attack and training for deep neural network based power quality disturbance classification. Engineering Applications of Artificial Intelligence. 2024. vol. 127.
90. Li W., Deka D., Wang R., Paternina M.R.A. Physics-Constrained Adversarial Training for Neural Networks in Stochastic Power Grids. IEEE Transactions on

Artificial Intelligence. 2024. vol. 5(3). pp. 1121–1131. DOI: 10.1109/TAI.2023.3236377.
91. Afrin A., Ardakanian O. Adversarial Attacks on Machine Learning-Based State Estimation in Power Distribution Systems. Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy '23). 2023. pp. 446–458.
92. Tian J., Wang B., Wang Z., Cao K., Li J., Ozay M. Joint Adversarial Example and False Data Injection Attacks for State Estimation in Power Systems. IEEE Transactions on Cybernetics. 2022. vol. 52(12). pp. 13699–13713.
93. Ali M.N., Amer M., Elsisi M. Reliable IoT Paradigm with Ensemble Machine Learning for Faults Diagnosis of Power Transformers Considering Adversarial Attacks. IEEE Transactions on Instrumentation and Measurement. 2023. vol. 72. pp. 1–13.
94. Kim B., Shi Y., Sagduyu Y.E., Erpek T., Ulukus S. Adversarial Attacks against Deep Learning Based Power Control in Wireless Communications. Proceedings of the IEEE Globecom Workshops (GC Wkshps). 2021. pp. 1–6.
95. Manoj B.R., Sadeghi M., Larsson E.G. Adversarial Attacks on Deep Learning Based Power Allocation in a Massive MIMO Network. Proceedings of the ICC – IEEE International Conference on Communications. 2021. pp. 1–6.
96. Manoj B.R., Sadeghi M., Larsson E.G. Downlink Power Allocation in Massive MIMO via Deep Learning: Adversarial Attacks and Training. IEEE Transactions on Cognitive Communications and Networking. 2022. vol. 8(2). pp. 707–719.
97. Santos P.M., Manoj B.R., Sadeghi M., Larsson E.G. Universal Adversarial Attacks on Neural Networks for Power Allocation in a Massive MIMO System. IEEE Wireless Communications Letters. 2022. vol. 11(1). pp. 67–71.
98. Shabbir A., Manzoor H.U., Ahmed R.A., Halim Z. Resilience of Federated Learning Against False Data Injection Attacks in Energy Forecasting. Proceedings of the 2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST). 2024. pp. 245–249.
99. Xu S., Yu L., Lin X. Robust Low-Overhead Control of DER Reactive Power under Adversarial Attacks and Uncertainty. Proceedings of the ICC 2024 – IEEE International Conference on Communications. 2024. pp. 3097–3103.
100. Kuzlu M., Sarp S., Catak F.O., Cali U., Zhao Y., Elma O., Guler O. Analysis of deceptive data attacks with adversarial machine learning for solar photovoltaic power generation forecasting. Electr. Eng. 2024. vol. 106(2). pp. 1815–1823.
101. Zideh M.J., Khalghani M.R., Solanki S.K. An unsupervised adversarial autoencoder for cyber attack detection in power distribution grids. Electric Power Systems Research. 2024. vol. 232.
102. Zeng L., Qiu D., Sun M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. Applied Energy. 2022. vol. 324.
103. Zografopoulos I., Hatziargyriou N.D., Konstantinou C. Distributed Energy Resources Cybersecurity Outlook: Vulnerabilities, Attacks, Impacts, and Mitigations. IEEE Systems Journal. 2023. vol. 17. pp. 6695–6709.
104. Santana E.J., Silva R.P., Zarpelão B.B., Barbon Junior S. Detecting and Mitigating Adversarial Examples in Regression Tasks: A Photovoltaic Power Generation Forecasting Case Study. Information. 2021. vol. 12(10).
105. Youssef E.-N.S., Labeau F., Kassouf M. Adversarial Dynamic Load-Altering Cyberattacks Against Peak Shaving Using Residential Electric Water Heaters. IEEE Transactions on Smart Grid. 2024. vol. 15(2). pp. 2073–2088.
106. Bhattacharjee A., Bai G., Tushar W., Verma A., Mishra S., Saha T.K. DeeBBAA: A Benchmark Deep Black-Box Adversarial Attack Against Cyber–Physical Power

Systems. IEEE Internet of Things Journal. 2024. vol. 11(24). pp. 40670–40688. DOI: 10.1109/JIOT.2024.3454257.
107. Genç D., Özuysal M., Tomur E. A Taxonomic Survey of Model Extraction Attacks. Proceedings of the 2023 IEEE International Conference on Cyber Security and Resilience (CSR). 2023. pp. 200–205.
108. Fan J., Yan Q., Li M., Qu G., Xiao Y. A Survey on Data Poisoning Attacks and Defenses. Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). 2022. pp. 48–55.
109. Wang S., Ko R.K.L., Bai G., Dong N., Choi T., Zhang Y. Evasion Attack and Defense on Machine Learning Models in Cyber-Physical Systems: A Survey. IEEE Communications Surveys & Tutorials. 2024. vol. 26. pp. 930–966.
110. El-Toukhy A.T., Mahmoud M., Bondok A.H., Fouda M.M., Alsabaan M. Evasion Attacks in Smart Power Grids: A Deep Reinforcement Learning Approach. Proceedings of the 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC). 2024. pp. 708–713.
111. Ali Alatwi H., Morisset C. Threat Modeling for Machine Learning-Based Network Intrusion Detection Systems. Proceedings of the IEEE International Conference on Big Data (Big Data). 2022. pp. 4226–4235.
112. Qiu S., Liu Q., Zhou S., Wu C. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied sciences. 2019. vol. 9(5).
113. Tuna O.F., Catak F.O., Eskil M.T. Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples. Multimed Tools Appl. 2022. vol. 81. pp. 11479–11500.
114. Dubrovkin J. Evaluation of undetectable perturbations of the peak parameters estimated by the least square curve fitting of analytical signal consisting of overlapping peaks. Chemom. and Intel. Lab. Syst. 2016. vol. 153. pp. 9–21.
115. Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 2002. vol. 86(11). pp. 2278–2324.
116. Liu X., Cheng M., Zhang H., Hsieh C.J. Towards Robust Neural Networks via Random Self-ensemble. Computer Vision – ECCV 2018. Lecture Notes in Computer Science. 2018. vol. 11211. pp. 381–397.
117. McCarthy A., Ghadafi E., Andriotis P., Legg P. Functionality-Preserving Adversarial Machine Learning for Robust Classification in Cybersecurity and Intrusion Detection Domains: A Survey. Journal of Cybersecurity and Privacy. 2022. vol. 2(1). pp. 154–190.
118. Goodfellow I., McDaniel P., Papernot N. Making Machine Learning Robust Against Adversarial Inputs. Communications of the ACM. 2018. vol. 61(7). pp. 56–66.
119. Li J., Yang Y., Sun J., Tomsovic K.L., Qi H. Towards Adversarial-Resilient Deep Neural Networks for False Data Injection Attack Detection in Power Grids. Proceedings of the 32nd International Conference on Computer Communications and Networks (ICCCN). 2023. pp. 1–10.
120. Li Y., Wang Y. Defense Against Adversarial Attacks in Deep Learning. Applied Sciences. 2018. vol. 9(1).
121. Zantedeschi V., Nicolae M.-I., Rawat A. Efficient Defenses Against Adversarial Attacks. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17). 2017. pp. 39–49.
122. Sarker A., Shen H., Sen T., Mendelson Q. Efficient Black-Box Adversarial Attacks for Deep Driving Maneuver Classification Models. Proceedings of the 18th International Conference on Mobile Ad Hoc and Smart Systems. 2021. pp. 536–544.
123. Gibert D., Zizzo G., Le Q., Planes J. Adversarial Robustness of Deep Learning-Based Malware Detectors via (De)Randomized Smoothing, IEEE Access. 2024. vol. 12. pp. 61152–61162.

124. Yuan X., He P., Zhu Q., Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. IEEE Transactions on Neural Networks and Learning Systems. 2019. vol. 30(9). pp. 2805–2824.
125. Ozturk E., Mesut A. Comparison of Learned Image Compression Methods and JPEG. Proceedings of the 2024 Innovations in Intelligent Systems and Applications Conference (ASYU). 2024. pp. 1–6.
126. Zhou Z., Firestone C. Humans can decipher adversarial images. Nat. Commun. 2019. vol. 10(1).
127. Mohammed A., Ali Z., Ahmad I. Enhancing adversarial robustness with randomized interlayer processing. Expert Systems with Applications. 2024. vol. 245.
128. Shi Y., Zeng H., Nguyen T.T. Adversarial Machine Learning for Network Security. Proceedings of the International Symposium on Technologies for Homeland Security. 2019. pp. 1–7.
129. Eleftheriadis C., Symeonidis A., Katsaros P. Adversarial robustness improvement for deep neural networks. Machine Vision and Applications. 2024. vol. 35(3).
130. Han S., Lin Ch., Shen Ch., Wang Q., Guan X. Interpreting Adversarial Examples in Deep Learning: A Review. ACM Comput. Surv. 2023. vol. 55. no. 14s. pp. 1–38.
131. Li B., Liu W. WAT: Improve the Worst-Class Robustness in Adversarial Training. In Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37(12). pp. 14982–14990.
132. Liu N., Du M., Guo R., Liu H., Hu X. Adversarial Attacks and Defenses: An Interpretation Perspective. SIGKDD Explor. Newsl. 2021. vol. 23(1). pp. 86–99.
133. Peng Y., Fu G., Luo Y., Yu Q., Wang L. CNN-based Steganalysis Detects Adversarial Steganography via Adversarial Training and Feature Squeezing. Proceedings of the 4th International Conference on Information Science, Parallel and Distributed Systems (ISPDS). 2023. pp. 165–169.
134. Wang W., Park Y., Lee T., Molloy I., Tang P., Xiong L. Utilizing Multimodal Feature Consistency to Detect Adversarial Examples on Clinical Summaries. Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics. 2020. pp. 259–268.
135. Shao J., Geng S., Fu Z., Xu W., Liu T., Hong S. CardioDefense: Defending against adversarial attack in ECG classification with adversarial distillation training. Biomedical Signal Processing and Control. 2024. vol. 91. DOI: 10.1016/j.bspc.2023.105922.
136. Li H., Xu X., Zhang X., Yang Sh., Li B. Qeba: Query-efficient boundary-based blackbox attack. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 1221–1230.
137. RyouD Ha I., Yoo H., Kim D., Han B. Robust Image Denoising Through Adversarial Frequency Mixup. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. pp. 2723–2732.
138. Tyukin I.Y., Higham D.J., Gorban A.N. On Adversarial Examples and Stealth Attacks in Artificial Intelligence Systems. Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2020. pp. 1–6.
139. Shimizu Y. Automatic Design System with Generative Adversarial Network and Vision Transformer for Efficiency Optimization of Interior Permanent Magnet Synchronous Motor. IEEE Transactions on Industrial Electronics. 2024. vol. 71(11). pp. 14600–14609.
140. Barth A., Rubinstein B.I.P., Sundararajan M., Mitchell J.C., Song D., Bartlett P.L. A Learning-Based Approach to Reactive Security. IEEE Transactions on Dependable and Secure Computing. 2012. vol. 9(4). pp. 482–493.

141. Aly A., Iqbal S., Youssef A., Mansour E. MEGR-APT: A Memory-Efficient APT Hunting System Based on Attack Representation Learning. IEEE Transactions on Information Forensics and Security. 2024. vol. 19. pp. 5257–5271.
142. Deng Y., Zheng X., Zhang T., Chen C., Lou G., Kim M. An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom). 2020. pp. 1–10.
143. He X., Yao J., Wang Y., Tang Z., Cheung K.C., See S., Han B., Chu X. NAS-LID: Efficient Neural Architecture Search with Local Intrinsic Dimension. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37(6). pp. 7839–7847.
144. Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18). 2018. pp. 2154–2156.
145. Kwatra S., Torra V. Data Reconstruction Attack against Principal Component Analysis. Security and Privacy in Social Networks and Big Data (SocialSec 2023). Lecture Notes in Computer Sciences. 2023. vol. 14097. pp. 79–92.
146. Pachika S., Reddy A.B., Pachika B., Karnam A. Generative Adversarial Networks: Overview. Proceedings of the Fifth International Conference on Computer and Communication Technologies (IC3T 2023). Lecture Notes in Computer Sciences. 2024. vol. 897. pp. 319–328.
147. Li H., Yu W., Huang H. Strengthening transferability of adversarial examples by adaptive inertia and amplitude spectrum dropout. Neural Networks. 2023. vol. 165. pp. 925–937.
148. Chen Z. Robust Sparse Online Learning through Adversarial Sparsity Constraints. Proceedings of the 2024 9th IEEE International Conference on Smart Cloud (SmartCloud). 2024. pp. 42–47.
149. Liu X., Chen X., Cheng J., Zhou L., Chen L., Li C., Zu S. Simulation of Complex Geological Architectures Based on Multistage Generative Adversarial Networks Integrating with Attention Mechanism and Spectral Normalization. IEEE Transactions on Geoscience and Remote Sensing. 2023. vol. 61. pp. 1–15.
150. Ahmed S., Islam S. Methods in detection of median filtering in digital images: a survey. Multimed. Tools Appl. 2023. vol. 82. pp. 43945–43965.
151. Elderman R., Pater L.J.J., Thie A.S., Drugan M.M., Wiering M.A. Adversarial Reinforcement Learning in a Cyber Security Simulation. Proceedings of the 9th International Conference on Agents and Artificial Intelligence. 2017. pp. 559–566.
152. Ling Y., Yong Z., Pengfei W. Euclidean and Rapid Jacobian-based Saliency Maps Attacks. Proceedings of the 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). 2021. pp. 355–361.
153. Vorobeychik Y., Kantarcioglu M. Introduction. Adversarial Machine Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2018. pp. 1–4.
154. Kotenko I., Saenko I., Lauta O., Vasiliev N., Sadovnikov V. A Noise-Based Approach Augmented with Neural Cleanse and JPEG Compression to Counter Adversarial Attacks against Image Classification Systems. Proceedings of the 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP). 2025. pp. 576–583.

Котенко Игорь Витальевич — д-р техн. наук, профессор, заслуженный деятель науки российской федерации, главный научный сотрудник, руководитель лаборатории, лаборатория проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: безопасность компьютерных сетей, включая управление политиками безопасности,

контроль доступа, аутентификацию, анализ безопасности, обнаружение компьютерных атак, брандмауэры, защиту от вирусов и сетевых червей, анализ и проверку протоколов безопасности и систем информационной безопасности, защиту программного обеспечения от взлома и управление цифровыми правами, технологии моделирования и визуализации для борьбы с кибертерроризмом. Число научных публикаций — 1000. ivkote@comsec.spb.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181; факс: +7(812)328-4450.

Саенко Игорь Борисович — д-р техн. наук, профессор, главный научный сотрудник, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: автоматизированные информационные системы, информационная безопасность, обработка и передача данных по каналам связи, теория моделирования и математическая статистика, теория информации. Число научных публикаций — 500. ibsaen@comsec.spb.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328–7181; факс: +7(812)328-4450.

Лаута Олег Сергеевич — д-р техн. наук, доцент, профессор кафедры, кафедра комплексного обеспечения информационной безопасности, Государственный университет морского и речного флота имени адмирала С.О. Макарова. Область научных интересов: защита от компьютерных атак. Число научных публикаций — 200. laos-82@yandex.ru; улица Двинская, 5/7, 198035, Санкт-Петербург, Россия; р.т.: +7(911)842-0228.

Садовников Владимир Евгеньевич — аспирант лаборатории, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: защита от компьютерных атак. Число научных публикаций — 25. bladimir1998@mail.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181; факс: +7(812)328-4450.

Ичетовкин Егор Андреевич — аспирант, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: защита от компьютерных атак. Число научных публикаций — 7. egor.email@list.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181; факс: +7(812)328-4450.

Ли Вэй — профессор, колледж компьютерных наук и технологий, Харбинский инженерный университет. Область научных интересов: защита от компьютерных атак, машинное обучение, искусственный интеллект. Число научных публикаций — 200. wei.li@hrbeu.edu.cn; Nantong St., Nangang District, 145, 150001, Харбин, Китай; р.т.: 0086(0451)8251921.

Поддержка исследований. Исследование частично поддержано бюджетной темой FFZF-2025-0016.