DOI 10.15622/ia.25.1.4

A. RAJEEV, RAVIRAJ P.

SPATIOTEMPORAL AND BEHAVIORAL FEATURE-AWARE MODEL WITH XCEPTIONCAPSULE FUSION FOR DEEPFAKE DETECTION

Rajeev A., Raviraj P. **Spatiotemporal and Behavioral Feature-Aware Model with XceptionCapsule Fusion for Deepfake Detection.**

Abstract. Deepfake detection continues to pose significant challenges, primarily because existing methods often suffer from key limitations, including reliance on individual frame analysis, vulnerability to low-resolution or compressed videos, and inability to capture temporal inconsistencies. Furthermore, traditional face detection techniques frequently fail under challenging conditions such as poor lighting or occlusion, while many models struggle with subtle manipulations due to inadequate feature extraction and overfitting on limited datasets. To address the drawbacks of existing deepfake detection approaches, this research proposes a Face and Motion-Aware Detection Framework that integrates both spatial and temporal information. The framework begins with a preprocessing stage that extracts video frames at a fixed rate to ensure temporal consistency. Facial regions and detailed landmarks are accurately detected using BlazeFace and MediaPipe Face Mesh. These features are then processed by the proposed XceptionCapsule Net, which combines the spatial feature extraction capabilities of the Xception model with the hierarchical and viewpoint-aware representation of Capsule Networks (CapsNet), and the temporal dependency modeling power of a Bidirectional Long Short-Term Memory (BiLSTM) layer. The architecture incorporates Global Average Pooling, Flatten, and fully connected layers, with Sigmoid activation for binary classification. Extensive evaluations on the FaceForensics++ (FF++) and Celeb-DF datasets demonstrate strong performance, achieving up to 99.31% accuracy and 99.99% Area Under the Curve (AUC). The results validate the framework's effectiveness, precision, and generalization across various video qualities and manipulation scenarios.

Keywords: deepfake detection, XceptionCapsule Net, Face Mesh, BlazeFace, facial landmark extraction, video forensics.

1. Introduction. The human face is one of the most defining features of an individual, making it a critical component in identity recognition systems [1]. As facial recognition technologies continue to evolve, so do concerns regarding their misuse, particularly with the rapid advancement of facial synthesis techniques. Among the most concerning developments is deepfake technology, an artificial intelligence-driven method that enables the seamless manipulation of facial features. This technology allows a person's face to be superimposed onto another's individual in a video without consent, raising serious ethical, legal, and security concerns. Deepfakes pose serious threats as they can be exploited to spread misrepresentation, steal identities, and carry out harmful activities [2, 3].

Deepfake content, typically in the form of synthetic videos and images, is widely disseminated across social media platforms. Although digital image manipulation has existed since the early days of photography [4], it traditionally required significant expertise and manual

effort, often involving software such as Adobe Photoshop. In contrast, the advent of Artificial intelligence (AI)-based tools has democratized content manipulation. With minimal technical knowledge, users can now generate highly realistic fake videos, making face-swapping and fabricated scenarios appear convincingly authentic [5 – 7].

Deepfake videos are typically generated using machine learning algorithms that digitally alter a subject's appearance by replacing it with that of another individual. These alterations are typically grouped into three main categories:

Head Puppetry: Synchronizes the head and shoulder movements of the target with those of a source individual.

Face Swapping: Replaces one person's face with another while maintaining the original expressions.

Lip-Syncing: Alters lip movements to match audio content not originally spoken by the subject [6].

While traditional computer graphics methods have been employed for similar purposes, modern deepfake generation predominantly relies on deep learning (DL) techniques, including generative adversarial networks (GANs) and autoencoders, which significantly enhance the realism of synthetic media and complicate detection efforts [7, 8]. In addition to these, diffusion-based generative models have recently gained prominence in content synthesis. Models such as Stable Diffusion [9], DALL-E [10], MidJourney [11], and Sora [12] can generate or manipulate highly realistic images and videos from simple textual prompts. By progressively denoising random noise into coherent outputs, these models offer unprecedented control and realism, making synthetic media creation more accessible and further intensifying the challenges of deepfake detection [13, 14]. Recent studies report a substantial increase in deepfake content across digital platforms, intensifying concerns around fraud, misinformation, and privacy violations. To address this challenge, leading organizations such as Defense Advanced Research Projects Agency (DARPA), Facebook, and Google have launched initiatives focused on developing advanced deepfake detection technologies [15, 16]. Numerous DL-based approaches have emerged to address this threat. Techniques involving Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and hybrid models have shown promise in identifying manipulated media. Additionally, Deep Neural Networks (DNNs) have shown effectiveness in detecting fake news and misleading social media content [17, 18].

However, significant research gaps persist. A major limitation of many current models is their reliance on individual frame analysis, which ignores temporal inconsistencies such as unnatural head movements and

facial expressions. This highlights the importance of temporal modeling through RNNs, LSTMs, or transformer-based architectures to better understand sequential patterns in videos. Additionally, existing systems struggle with low-resolution or compressed videos, where common artifacts are masked, hindering manipulation detection. Face detection, a crucial preprocessing step, is especially vulnerable under real-world conditions, such as occlusion, poor lighting, and low resolution, often resulting in missed or inaccurate face localization. Moreover, as deepfake generation grows more sophisticated, many subtle alterations bypass detection due to the limited feature extraction capabilities of traditional models. Another pressing issue is overfitting, primarily driven by scarce and homogeneous training datasets, which undermines model generalization across diverse manipulation types and datasets. These challenges underscore the urgent need for a versatile, temporally aware, and generalizable deepfake detection framework that can operate effectively under a variety of media conditions and evolving manipulation techniques. To bridge these gaps, this study presents a face and motion-aware framework that effectively utilizes spatial and temporal information to achieve reliable and high-accuracy deepfake detection. The primary contributions are as follows:

1. An innovative framework designed to improve the accuracy and reliability of deepfake detection by combining spatial and temporal feature analysis.
2. The framework ensures consistent frame extraction at a fixed sampling rate to maintain video-level coherence. It utilizes BlazeFace [19] for efficient face localization and landmark detection, along with MediaPipe Face Mesh [20], to provide accurate spatial inputs under challenging conditions such as low resolution, occlusion, and lighting variations.
3. A hybrid deep neural network (DNN) is designed by integrating the Xception model [21] with Capsule Networks (CapsNet) [22] to capture spatial hierarchies and viewpoint variations. This architecture enhances sensitivity to fine-grained facial anomalies and improves generalization across different manipulation types.
4. Temporal dependency modeling is introduced through a Bidirectional Long Short-Term Memory (BiLSTM) layer, which captures sequential relationships and motion continuity between consecutive frames. This addition allows the model to detect temporal inconsistencies in blinking, lip movement, and head motion that are commonly present in manipulated videos.
5. The model employs Global Average Pooling (GAP), followed by Flatten and fully connected layers, with sigmoid activation to ensure robust binary classification.

This article is structured to offer a thorough and organized explanation of the proposed deepfake detection approach. Section 2 reviews relevant literature, outlining existing methods, notable contributions, and existing research gaps in deepfake detection. Section 3 explains the proposed approach in detail, covering the preprocessing steps, extraction of spatial and temporal features, and the overall architecture of the detection model. Section 4 presents the experimental setup and results, accompanied by a detailed performance evaluation using standard benchmark datasets. Finally, Section 5 summarizes the main outcomes and discusses potential future improvements to enhance the effectiveness, adaptability, and reliability of deepfake detection systems.

2. Literature Review. The rapid progress in deepfake generation technologies in recent years has raised serious concerns about the credibility and authenticity of digital content. In response, a range of DL-based detection approaches have been proposed, aiming to accurately differentiate authentic content from manipulated media. Existing literature explores various strategies, including hybrid models, attention mechanisms, and temporal analysis, to address the limitations of earlier frame-level or spatial-only methods. These efforts underscore the urgent need for robust, generalizable frameworks capable of detecting subtle visual and temporal inconsistencies across diverse datasets and manipulation techniques.

In paper [23] the authors introduced a novel approach to deepfake detection by analyzing the contribution of distinct facial regions using face cutout techniques. This study was included in our review because it addresses the overfitting problem in deepfake datasets and aligns with the emerging trend of leveraging facial region importance for more robust detection. Input images were augmented by selectively removing facial areas based on landmarks, creating diverse training samples, while training was conducted with an 80/10/10 split for training, validation, and testing. The methodology was evaluated on FF++ and Celeb-DFv2 datasets, achieving up to 91% accuracy on Celeb-DF and demonstrating the significance of external facial features, particularly the eyes, for detection. This work highlights a broader trend in deepfake research toward targeted data augmentation strategies and feature-aware model training, though it suggests further exploration is needed for video-based deepfake detection and cross-dataset generalization.

Paper [24] proposed a Frequency-Enhanced Self-Blended Images (FSBI) approach for deepfake detection, which blends images with themselves to introduce generic forgery artifacts and uses Discrete Wavelet Transform (DWT) to extract discriminative frequency-domain features. This study was included in our review because it demonstrates a trend

toward frequency-aware and artifact-generalized detection strategies that improve model robustness and cross-dataset generalization. The model was evaluated on FF++ and Celeb-DF datasets using both within-dataset and cross-dataset protocols, achieving an AUC of 95.49% when trained on FF++ and tested on Celeb-DF, highlighting strong adaptability to unseen manipulations. FSBI effectively mitigates overfitting to dataset-specific artifacts, and ablation studies confirmed the benefits of self-blending and frequency feature extraction. This work illustrates the broader research trend of combining spatial and frequency-domain analysis for deepfake detection, though performance remains lower for certain complex manipulations such as NeuralTextures (NT) and Face2Face (F2F), suggesting areas for further improvement in real-world scenarios.

In paper [25] the authors proposed an efficient deepfake detection framework using a hybrid ResNet-Swish-BiLSTM network, which combines residual learning for spatial features with recurrent modeling of temporal dependencies. The model was extensively tested on the FaceForensics++ (FF++) and Deepfake Detection Challenge (DFDC) datasets, achieving 96.23% accuracy on FF++ and 78.33% accuracy when aggregating FF++ and DFDC records. To assess generalization, they further conducted cross-corpus experiments using Celeb-DF, where performance dropped to AUC values of 71.56% (DFDC) and 70.04% (Celeb-DF) when trained on FF++, and 70.12% (FF++) and 65.23% (Celeb-DF) when trained on DFDC. This evaluation highlights a broader trend in deepfake detection research: models often perform well in database-internal evaluations but degrade under cross-dataset conditions due to differences in compression, resolution, and manipulation artifacts. The study was included in our review as it illustrates both the strength of hybrid spatial-temporal learning and the persistent challenge of cross-dataset generalization, which directly relates to our proposed approach.

Paper [26] introduced a deepfake detection framework, BMNet, that integrates BiLSTM to capture temporal dependencies across frames and multi-head self-attention (MHSA) to extract localized forgery features from facial regions. Unlike static CNN-based methods, BMNet explicitly models both temporal and regional inconsistencies, addressing a key gap in prior research. The model was evaluated on four benchmark datasets (FF++, UADFV, Celeb-DF, DFDC) and achieved 95.54%, 92.18%, 80.20%, and 84.72% accuracy, respectively, demonstrating consistent improvements across varying data sources. Importantly, the inclusion of landmark-based features enhanced interpretability and robustness, with ablation studies confirming the contribution of both BiLSTM and MHSA modules. This work was included in our survey because it exemplifies a spatiotemporal

trend in deepfake detection and shows stronger cross-dataset adaptability than handcrafted or purely spatial CNN models, while still highlighting the persistent performance gap on Celeb-DF compared to easier datasets.

In [27] the authors introduced a stacking-based ensemble framework that fuses deep features from Xception and EfficientNet-B7, followed by feature ranking and classification using an MLP meta-learner. The model achieved 96.33% accuracy on Celeb-DF (V2) and 98.00% on FaceForensics++ (FF++), outperforming the individual base models. Their experimental setup used clear train/validation/test splits for both datasets, ensuring fair evaluation and reproducibility. This work was selected as it represents the ensemble and meta-learning trend in deepfake detection, focusing on combining complementary deep models with feature selection for improved robustness. However, the authors also noted limitations, including increased computational cost, reduced interpretability, and potential overfitting risks, issues that remain common across ensemble approaches. The study is significant because it demonstrates how carefully designed stacking can improve cross-dataset generalization, while also highlighting challenges in balancing accuracy with efficiency and transparency.

Paper [28] introduced a self-supervised BEiT-HPR (Hierarchical PatchReducer) model for efficient facial deepfake detection aimed at addressing the high computational cost associated with existing detection systems. The study was motivated by the growing difficulty of distinguishing real from fake facial videos due to the rapid advancement of generative models and the impracticality of deploying complex models in resource-limited environments. Experimental results across FF++, Celeb-DF, and DFD datasets demonstrated notable efficiency gains and strong detection performance. However, the approach may face limitations in handling unseen forgery types or cross-dataset generalization due to its reliance on self-supervised pretraining within limited data domains.

In paper [29] the authors proposed ConvNext-PNet, an interpretable and explainable deep learning framework for detecting visual deepfakes, aiming to enhance trust and transparency in AI-based forgery detection. The motivation behind this work was to address the limitation of existing deepfake detection models that, despite achieving high accuracy, lack interpretability and fail to justify their classification decisions. By integrating prototype learning into a modified ConvNext architecture, the model not only improves discriminative feature learning but also provides visual reasoning for its predictions, thereby increasing user trust. Experimental evaluations on benchmark datasets such as FF++, CelebDF, DFDC-P, and DFF demonstrated high robustness and effective detection of

visual manipulations. However, the added interpretability components may introduce additional computational overhead, potentially limiting real-time applicability in large-scale or streaming scenarios.

In [30] the authors introduced an improved deepfake video detection framework based on a Convolutional Vision Transformer (CViT2) that combines the strengths of CNNs and Vision Transformers to enhance detection accuracy and generalization. The model was proposed to address the limitations of existing CNN-based methods, which often struggle to capture global dependencies and contextual information in video frames. By employing a CNN for extracting learnable spatial features and a Vision Transformer for modeling long-range relationships using attention mechanisms, the CViT2 architecture effectively identifies subtle manipulation cues in deepfake videos. Experiments conducted across multiple benchmark datasets, including DFDC, FF++, Celeb-DF v2, and DeepfakeTIMIT, demonstrated high accuracy and strong cross-dataset robustness. However, the model's complex structure and heavy training requirements may restrict its real-time applicability and scalability in low-resource environments.

Paper [31] proposed a robust face forgery detection framework that integrates both local and global texture information to enhance detection accuracy and generalization. The method was introduced to address the limitations of existing CNN-based approaches, which often overfit training data and fail to capture subtle forgery traces across diverse sources and post-processing variations. By employing a two-stream architecture combining RGB and texture features, along with an adaptive feature fusion and attention mechanism, the model effectively exposes fine-grained artifacts at multiple scales. This approach improves robustness and feature discrimination, leading to better performance across benchmark datasets. However, the model's complexity and computational cost may limit its deployment in real-time or resource-constrained environments.

Paper [32] proposed a hybrid CNN-LSTM model for video deepfake detection that leverages optical flow features to capture both spatial and temporal cues from video frames. The approach was introduced to overcome the limitation of conventional CNN-based methods, which primarily focus on spatial information and fail to exploit temporal dependencies between frames. This hybrid strategy enhances detection accuracy even with a limited number of samples, achieving competitive results on benchmark datasets such as DFDC, FF++, and Celeb-DF. However, the model's performance may still be constrained by moderate accuracy on challenging datasets and dependence on precise optical flow computation, which can increase computational overhead.

These studies highlight the evolving sophistication of deepfake detection frameworks and the need to incorporate spatial, temporal, and attention-based mechanisms. Nonetheless, existing methods often fail to address issues such as generalization [27] to unseen manipulations, real-world distortions, and computational efficiency. Table 1 offers a detailed overview of the main insights and approaches highlighted in the literature review.

Table 1. Overview of the literature review

No.	Methodology	Key Features	Dataset & Split	Performance	Limitations / Observations
[23]	Face cutout-based augmentation	Region-specific face removal; data augmentation to reduce overfitting	FF++, Celeb-DFv2; 80/10/10 train/val/test	91% on Celeb-DF	Eyes most impactful; limited exploration for video sequences & cross-dataset validation
[24]	FSBI (Frequency Enhanced Self-Blended Images) + DWT	Self-blended images to create generic artifacts; frequency domain features	FF++, Celeb-DF; within- and cross-dataset	AUC 95.49% (FF++→Celeb-DF)	Lower performance on complex manipulations (F2F, NT); needs real-world evaluation
[25]	ResNet-Swish-BiLSTM hybrid	Residual CNN for spatial + BiLSTM for temporal modeling	FF++, DFDC, Celeb-DF; cross-corpus evaluation	96.23% (FF++), 78.33% (FF+++DFDC); AUC 70-71% cross-dataset	Strong intra-dataset performance, weak cross-dataset generalization
[26]	BMNet (BiLSTM + MHSA)	Temporal + regional attention; landmark-based features	FF++, UADFV, Celeb-DF, DFDC	95.54%, 92.18%, 80.20%, 84.72%	Cross-dataset gap on Celeb-DF; highlights spatiotemporal trend
[27]	Stacking ensemble (Xception + EfficientNet-B7 + MLP)	Feature fusion + meta-learning; deep feature ranking	FF++, Celeb-DFv2; clear train/val/test split	98% (FF++), 96.33% (Celeb-DFv2)	Increased computation; interpretability issues; risk of overfitting

Continuation of the Table 1

No.	Methodology	Key Features	Dataset & Split	Performance	Limitations / Observations
[28]	Self-Supervised BEiT-HPR (Hierarchical PatchReducer)	Hierarchical patch reduction to lower computation; self-supervised BEiT backbone	FF++, Celeb-DF, DFD; standard benchmark split	83.92% (FF++), 97.59% (Celeb-DF), 98.25% (DFD)	Strong accuracy with reduced complexity; may face limitations on unseen forgery types
[29]	ConvNext-PNet (Prototype-based ConvNext)	Prototype learning for interpretability; explainable visual reasoning	FF++, CelebDF, DFDC-P, DFF	High robustness across datasets (specific accuracy not stated)	Provides interpretability; adds computational overhead affecting real-time use
[30]	Convolutional Vision Transformer (CViT2)	CNN for spatial extraction + ViT for global context with attention	DFDC, FF++, Celeb-DFv2, DeepfakeTIMIT, TrustedMedia	95% (DFDC), 94.8% (FF++), 98.3% (Celeb-DFv2), 76.7% (TIMIT)	Excellent cross-dataset accuracy; high model complexity and training cost
[31]	Two-stream texture-aware network	Integration of global and local texture features; adaptive fusion and attention modules	Multiple benchmark datasets (FF++, Celeb-DF, etc.)	Outperformed recent SOTA methods	High computational complexity; limited real-time applicability
[32]	Hybrid CNN-LSTM with optical flow	Temporal feature extraction via optical flow; hybrid spatial-temporal learning	DFDC, FF++, Celeb-DF; ≤100 frames/sample	66.26% (DFDC), 91.21% (FF++), 79.49% (Celeb-DF)	Moderate accuracy on complex datasets; relies on precise optical flow computation

A comprehensive review of recent studies reveals several key trends and research directions in the domain of deepfake detection. Contemporary approaches increasingly emphasize spatial-temporal integration, where models such as those in [25 – 27, 30, 32] combine spatial and temporal features to overcome the limitations of static frame analysis. Techniques incorporating BiLSTM or hybrid CNN-LSTM frameworks have demonstrated the ability to capture motion inconsistencies across consecutive frames, thereby improving temporal awareness and detection accuracy. Another significant development is the adoption of attention-based and

transformer-based mechanisms, as observed in BMNet [26], CViT2 [30], and GRAM [31], which capture global relationships and multi-scale feature dependencies to achieve better generalization across diverse datasets. Furthermore, frequency and artifact-based analysis methods, including DWT and FSBI [24], have proven effective in identifying high-frequency inconsistencies that are often imperceptible in RGB domains. Researchers have also explored ensemble and meta-learning strategies [27, 31], where complementary features from CNN and transformer architectures are fused to enhance stability and robustness, albeit at increased computational cost. In parallel, there is growing attention toward explainable and efficient detection models [28, 29], which ensure interpretability and real-time applicability – key requirements for deployment in social media monitoring and law enforcement systems. Despite these advancements, a persistent cross-dataset generalization gap remains a major challenge, as many models still exhibit significant performance degradation when evaluated on unseen data [25, 27]. Overall, the literature reflects a clear evolution from static, handcrafted models toward dynamic, spatial-temporal, and attention-driven architectures, with a heightened emphasis on explainability, efficiency, and adaptability. Building on these emerging trends, the proposed research introduces a unified Face and Motion-Aware XceptionCapsule Net that synergistically integrates spatial and temporal cues to achieve enhanced robustness and reliability in detecting diverse deepfake manipulations.

3. Proposed methodology. Existing deepfake detection models encounter several critical challenges that significantly impact their accuracy and robustness [23, 24]. A primary limitation is their reliance on individual frame analysis, which performs inadequately on low-resolution or highly compressed videos where visual artifacts such as blurring and pixelation are less perceptible [25, 26]. This diminishes the model's effectiveness in identifying manipulated content. Furthermore, face detection in such models is highly sensitive to environmental factors, including poor lighting, occlusions, and extreme facial angles, often resulting in missed or inaccurate face localizations [27, 28]. Another key limitation lies in the disregard of temporal inconsistencies. Since many models process frames independently, they fail to capture unnatural head movements or inconsistent facial expressions that may indicate tampering [29, 30]. As deepfake generation techniques become increasingly sophisticated, subtle manipulations with minimal visual cues often go undetected due to inadequate feature extraction mechanisms [21]. Additionally, many models suffer from overfitting, primarily caused by limited and homogeneous training data, thereby restricting their generalization to unseen deepfake formats and real-world media conditions [33, 34].

To overcome these drawbacks, this work proposes an innovative Face and Motion-Aware Detection Framework. The preprocessing step enhances temporal consistency by extracting video frames at a uniform frame rate. For face detection, the framework employs BlazeFace, which enables efficient and accurate facial region extraction. Subsequently, MediaPipe Face Mesh is used to extract detailed 3D facial landmarks, improving the granularity of spatial feature identification. The extracted features are then fed into the proposed XceptionCapsule-BiLSTM Net, which integrates the Xception network, CapsNet, and a Bidirectional Long Short-Term Memory (BiLSTM) layer to jointly capture spatial details and temporal dependencies. The Xception module, utilizing depthwise separable convolutions, is effective at extracting fine-grained spatial features that are critical for identifying deepfake artifacts. In parallel, the CapsNet component preserves spatial hierarchies and models the interrelationships between features, while the BiLSTM layer captures sequential frame relationships and motion dynamics, enabling the detection of temporal anomalies such as unnatural motion or facial distortions. The model architecture incorporates Global Average Pooling (GAP) to reduce dimensionality, a Flatten layer to vectorize the features, and Fully Connected layers to extract high-level representations. A Sigmoid activation function is used for binary classification between real and fake content. The strength of this approach lies in its unified spatial-temporal modeling via the XceptionCapsule-BiLSTM Net, which enhances the detection of subtle manipulations and improves robustness against common real-world challenges such as low resolution, compression artifacts, and sophisticated forgery techniques. Experimental evaluations (see Section 4) confirm the efficacy and generalizability of the proposed approach on multiple benchmark datasets. Figure 1 illustrates the block diagram of the proposed deepfake detection methodology.

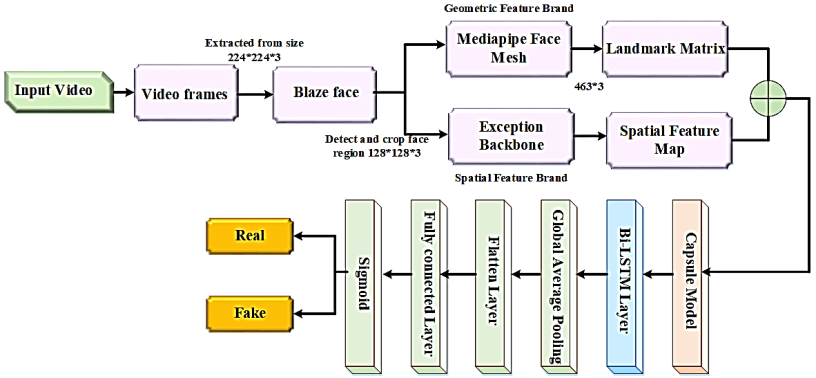


Fig. 1. Block diagram of the proposed methodology

3.1. Preprocessing: BlazeFace Mesh Processor. The preprocessing stage plays a crucial role in enhancing deepfake detection, especially when dealing with low-resolution or highly compressed video inputs. Such conditions often obscure subtle artifacts and degrade the accuracy of face detection, particularly under poor lighting, occlusions, or low image quality. To overcome these drawbacks, this work introduces a novel BlazeFace Mesh Processor, designed to ensure robust and consistent facial region extraction. Initially, input videos are decomposed into individual frames at a uniform frame rate, thereby preserving temporal coherence essential for detecting sequential anomalies. The BlazeFace model is then employed to perform fast and efficient face detection, generating precise bounding boxes for each detected face within the frames. Subsequently, MediaPipe Face Mesh is applied to these detected faces to extract dense 3D facial landmarks. This step significantly enhances the granularity of spatial feature representation, improving the identification of subtle facial deformations introduced by deepfake manipulations. The integration of BlazeFace and MediaPipe in a unified pipeline ensures both speed and precision in preprocessing, laying a strong foundation for effective feature extraction and subsequent deepfake detection.

3.1.1. Video Frames. Video frames are the basic units for analyzing spatial and temporal features in deepfake detection. A video is first converted into individual frames at a consistent frame rate (typically 30 fps) to preserve temporal coherence, allowing the framework to detect anomalies such as unnatural facial expressions, inconsistent head movements, or temporal flickering. Each frame is then processed using the BlazeFace model, which accurately localizes faces by generating precise bounding boxes, providing a reliable foundation for subsequent manipulation detection.

3.1.2. BlazeFace. To accurately isolate face regions from individual video frames, the proposed framework employs BlazeFace, a real-time neural network-based face detector optimized for mobile and embedded Graphics Processing Units (GPUs). It is designed to deliver high inference speeds while maintaining accurate face localization and keypoint estimation. This makes it well-suited for handling extensive video datasets in deepfake detection applications.

BlazeFace uses depthwise separable convolution with 5×5 kernels to achieve a larger receptive field at low computational cost. The network architecture is built using BlazeBlocks and Double BlazeBlocks, which preserve spatial resolution while minimizing computational overhead. The model operates on input frames of size $128 \times 128 \times 3$, and outputs both

bounding boxes and six key facial landmarks, including positions of the eyes, nose, ears, and mouth center. Specifically, the output tensors are:

- **Bounding boxes**: [N, 4], where N is the number of faces detected, and 4 represents the coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$ of each face.
- **Facial landmarks**: [N, 6, 2], where 6 corresponds to the six keypoints and 2 represents (x, y) coordinates normalized relative to the input frame.

Figure 2 illustrates the structure of the BlazeBlock and Double BlazeBlock components.

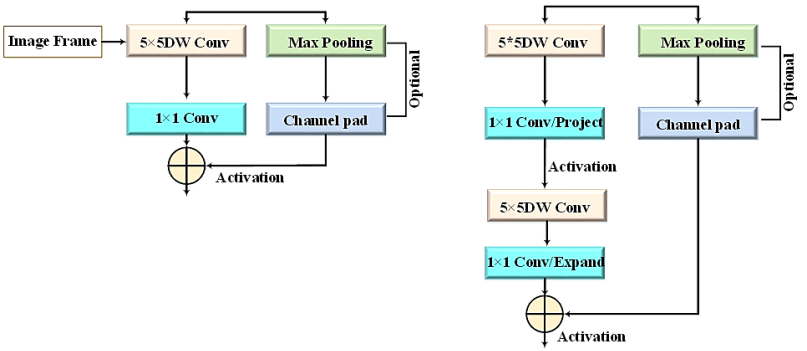


Fig. 2. Structure of BlazeBlock and Double BlazeBlock

Let $I \in \mathbb{R}^{H \times W \times 3}$ denote an input video frame. The model generates a set of anchor boxes $\{a_k\}_{k=1}^{K}$. For each anchor, the network predicts:

Bounding box offsets:

$$(\triangle x_k, \triangle y_k, \triangle w_k, \triangle h_k), \tag{1}$$

where $\triangle x_k$ denotes the horizontal offset of the bounding box center for anchor k; $\triangle y_k$ denotes the vertical offset of the bounding box center for anchor k; $\triangle w_k$ denotes the logarithmic scaling factor for the width of bounding box k; $\triangle h_k$ denotes logarithmic scaling factor for the height of bounding box k; k represents the index of the anchor box among total anchors.

Keypoint coordinates:

$$\{P_{ki}\}_{i=1}^{6} \in \mathbb{R}^2, \tag{2}$$

where P_{ki} denotes the 2D coordinate (x, y) of the i^{th} facial landmark for anchor k; and $\mathbb{R}^2$ indicates each key point represented in 2D space.

The final bounding box b_k is computed as:

$$x'_k = x^a_k + \triangle x_k . w^a_{k,}\ y'_k = y^a_k + \triangle y_k . h^a_k\,, \tag{3}$$

where x'_k and y'_k denote the refined center coordinates of bounding box k; x^a_k and y^a_k denote the anchor box center coordinates for anchor k; $w^a_{k,}$ and $h^a_{k,}$ denote the anchor box width and height; and $\triangle x_k$ and $\triangle y_k$ denote the predicted offsets.

$$w'_k = w^a_k . \exp(\triangle w_k)\,, \qquad h'_k = h^a_k . \exp(\triangle h_k), \tag{4}$$

where w'_k and h'_k denote the refined width and height of bounding box k; w^a_k, and h^a_k denote the anchor width and height; $\triangle w_k$ and $\triangle h_k$ predicted scaling factors; and $\exp(.)$denotes the exponential function to ensure positive dimensions.

The facial keypoints are then computed as:

$$P'_{ki} = P_{ki} . (w'_k, h'_k) + (x'_k, y'_k), \tag{5}$$

where P'_{ki} denotes the refined location of landmark i for anchor k; P_{ki} represent the normalized keypoint coordinates; (w'_k, h'_k) represent the bounding box size; and (x'_k, y'_k) represent the bounding box center.

To reduce temporal jitter across frames, a common artifact in deepfake content, BlazeFace implements a weighted blending strategy in place of traditional Non-Maximum Suppression (NMS). The final refined bounding box is computed as:

$$b_{final} = \frac{\sum_{k \in N} w_k . b_k}{\sum_{k \in N} w_k}, \tag{6}$$

where w_k is the confidence score; b_{final} denotes the final bounding box after blending; and N is the set of overlapping anchor boxes. This enhances frame-to-frame consistency, helping the model distinguish genuine manipulations from jitter-induced artifacts, thereby increasing the reliability of detection.

3.1.3. MediaPipe Face Mesh. After detecting the face regions using BlazeFace, the framework integrates MediaPipe Face Mesh to extract a dense set of 3D facial landmarks. This enables the detection of unnatural

facial geometry and behavioral inconsistencies, which are often indicative of deepfake manipulations.

MediaPipe Face Mesh predicts a total of 468 3D facial landmarks, represented as:

$$L = \{(x_i, y_i, z)\}_{i=1}^{468} \in \mathbb{R}^{468\times 3}, \tag{7}$$

where L denotes the set of 3D landmark coordinates; and x_i, y_i, z denote the coordinates of landmark i.

These landmarks span the entire facial geometry, including contours, lips, nose, eyes, eyebrows, jawline, and forehead. The input to MediaPipe Face Mesh is the face crop provided by BlazeFace, resized and normalized to a tensor of size 128×128×3, consistent with the BlazeFace output bounding box region. This normalization removes scale and rotation variances, ensuring consistent landmark extraction across frames. The extracted landmarks are utilized to compute both geometric and behavioral features, which are highly sensitive to subtle facial manipulations introduced by generative adversarial networks (GANs). Several key features derived from these landmarks include:

Eye Aspect Ratio (EAR) for blink analysis:

$$EAR = \frac{\left||L_2 - L_6|\right| + ||L_3 - L_5||}{2.||L_1 - L_4||}. \tag{8}$$

This ratio helps monitor blinking behavior and can reveal synthetic inconsistencies.

Mouth Aspect Ratio (MAR) is used to identify lip-sync anomalies, which indicative of poor audio-visual alignment in deepfakes.

Facial symmetry deviation is calculated by comparing corresponding landmarks on the left and right sides of the face to assess asymmetry introduced by manipulation.

Temporal displacement of keypoints, which captures jitter or abnormal warping across consecutive frames, is a common artifact in manipulated videos.

The combined BlazeFace + MediaPipe Face Mesh pipeline establishes a robust preprocessing foundation by ensuring precise face localization and dense landmark tracking, even under adverse conditions such as low lighting, occlusion, or blurred resolution. By analyzing temporal coherence across frames, the framework detects subtle geometric distortions, expression inconsistencies, and behavioral anomalies, traits

rarely found in authentic content. Furthermore, it supports high-resolution tracking of facial dynamics, such as eye blinking, lip movement, and overall facial expression behavior, thereby improving deepfake detection reliability.

3.2. XceptionCapsule-BiLSTM Net. Existing deepfake detection models often analyze individual frames, missing temporal cues such as head movements or inconsistent facial expressions, which limits their performance on subtle manipulations. To address this, we propose the XceptionCapsule-BiLSTM Net, a hybrid architecture that processes sequences of N consecutive frames (16 in our experiments) to capture both spatial and temporal dependencies such as eye blinks, lip-syncing, and head motion. The model integrates Xception for high-resolution spatial feature extraction and CapsNet to preserve spatial hierarchies and inter-feature relationships. To further enhance temporal modeling, a BiLSTM layer is incorporated after the Capsule layer. The BiLSTM learns sequential dependencies across consecutive frame embeddings, enabling the network to recognize motion continuity and detect temporal inconsistencies such as abrupt facial transitions or unnatural movements. This addition ensures that the model processes video segments as coherent temporal sequences rather than isolated frames. Key layers include GAP for feature reduction, Flatten and Fully Connected layers for high-level representation, and a Sigmoid activation for binary classification. By combining spatial hierarchy learning (Xception + CapsNet) with temporal dependency modeling (BiLSTM), the proposed architecture significantly enhances robustness, accuracy, and generalization in detecting both obvious and subtle deepfakes in real-world scenarios.

3.2.1. Xception Model. The Xception model is particularly effective for deepfake detection due to its superior capability in spatial feature extraction. Unlike conventional CNNs, which perform standard convolution over both spatial and channel dimensions simultaneously, Xception employs depthwise separable convolutions, dividing the process into two stages: depthwise convolution individually processes each input channel using separate filters, while pointwise convolution uses 1×1 convolutions to merge the results from the depthwise operation. This decomposition results in significant computational efficiency and a reduction in the number of learnable parameters, while still maintaining the capacity to capture essential visual features. In the proposed framework, Xception processes RGB face crops extracted by BlazeFace, with an input tensor of size $128 \times 128 \times 3$. The network outputs feature maps of shape [batch_size, h, w, c], where h, w, and c correspond to the height, width, and number of channels of the final convolutional layer. These feature maps retain rich spatial information, enabling the detection of subtle artifacts introduced by

generative models, such as slight geometric distortions, texture inconsistencies, and skin tone mismatches.

The model's enhanced spatial sensitivity enables it to effectively distinguish authentic faces from manipulated ones, even in high-resolution or minimally altered fake videos. The architectural pipeline of the Xception model is illustrated in Figure 3, showcasing its layered depthwise separable convolutional structure optimized for detailed pattern recognition in facial analysis.

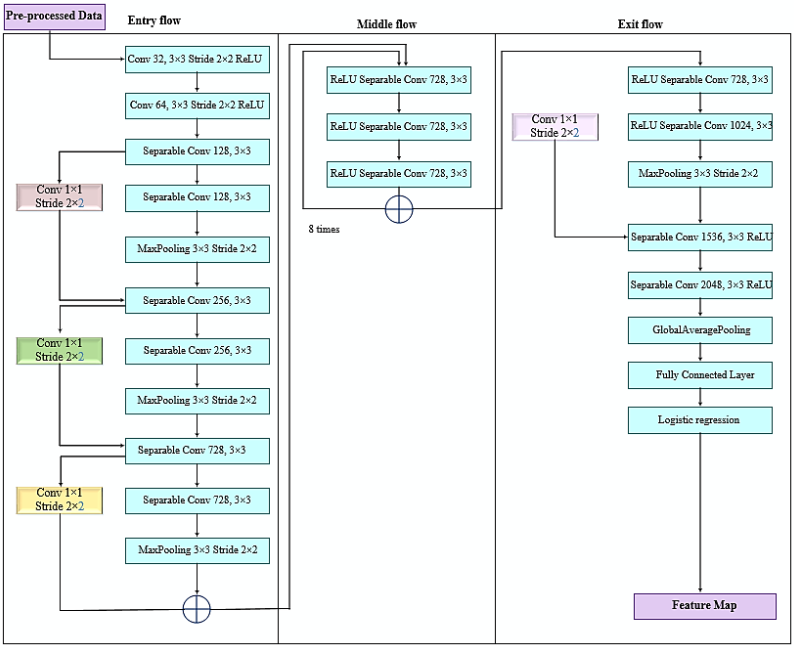


Fig. 3. Architecture of the Xception model

3.2.2. Capsule Model. The CapsNet is designed to overcome the limitations of traditional CNNs, particularly their inability to preserve spatial hierarchies and capture part-whole relationships in visual data. Unlike CNNs, which use scalar activations and pooling operations that often discard spatial information, CapsNet represents features as vectors, thereby encoding both the presence and pose (e.g., position, orientation, deformation) of detected entities.

The core building block of CapsNet is the capsule, a group of neurons whose activity vector $u_i \in \mathbb{R}^d$ encodes the instantiation parameters

of a specific object or object part. Capsules are organized hierarchically such that higher-level capsules model increasingly complex entities by aggregating information from lower-level capsules.

To support this hierarchy, CapsNet introduces a dynamic routing mechanism between layers. Each capsule i in the primary capsule layer predicts the output of capsule j in the next layer using a learned transformation matrix W_{ij}. The predicted output is computed as:

$$\hat{u}_{j|i} = W_{ij} u_i. \tag{9}$$

Here, $\hat{u}_{j|i}$ is the prediction vector from capsule i to capsule j, and u_i is the output vector of capsule i. These predictions are aggregated into a total input S_j for capsule j through a weighted sum of the predictions:

$$S_j = \sum_i c_{ij} \hat{u}_{j|i}. \tag{10}$$

Here, c_{ij} are the coupling coefficients that determine the contribution of capsule i to capsule j, and $\hat{u}_{j|i}$ is the predicted output from capsule i. These coefficients are computed through a routing-by-agreement mechanism, using a softmax function over initial logits b_{ij}, which are iteratively updated:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}. \tag{11}$$

The final output of capsule j, denoted v_j, is computed by applying a nonlinear function to s_j. This function ensures that short vectors are shrunk to nearly zero length (representing absence), while longer vectors are scaled to a length slightly below 1 (representing its presence), without affecting their orientation:

$$v_j = \frac{\| s_j \|^2}{1+\| s_j \|^2} \cdot \frac{s_j}{\| s_j \|^2}, \tag{12}$$

where the length $\| v_j \|$ represents the probability that the entity modeled by capsule j exists, and the direction of v_j encodes its pose and $\| s_j \|$ is the magnitude of input vector. This structure enables CapsNet to retain spatial hierarchies and model complex visual relationships, making it particularly

effective for detecting subtle, hierarchical inconsistencies inherent in deepfake content. The architecture of the CapsNet is illustrated in Figure 4.

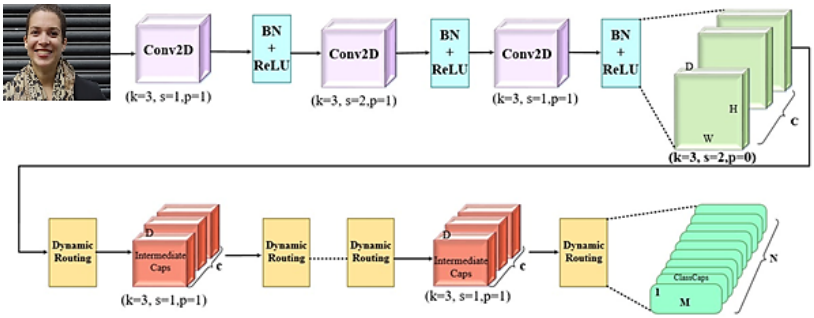


Fig. 4. Structure of CapsNet

This mechanism enables CapsNet to retain detailed spatial relationships and generalize effectively across transformations such as rotation and scaling. Consequently, CapsNet proves particularly effective in tasks requiring precise localization, pose estimation, and structure-preserving representations, such as facial analysis, medical imaging, and deepfake detection. The proposed XceptionCapsule Net architecture integrates the powerful spatial feature extraction capabilities of the Xception network with the hierarchical modeling strength of CapsNet. By leveraging key architectural components such as Global Average Pooling, Flatten, Dense layers, Capsule layers, and a final Sigmoid activation function, the model robustly captures subtle visual artifacts and spatial dependencies. This well-structured design significantly improves the precision and reliability of distinguishing between real and fake videos.

3.2.3. BiLSTM Layer. To capture temporal dependencies across consecutive video frames, a BiLSTM layer is introduced after the Capsule layer. While the Capsule Network effectively models spatial hierarchies and part–whole relationships, it does not account for sequential variations in motion or expression. The BiLSTM addresses this by processing sequential capsule feature embeddings from consecutive frames, learning both forward and backward temporal correlations. This enables the model to capture motion dynamics such as eye blinking, lip movement, and head rotation over time, patterns that are often inconsistent in manipulated content. Mathematically, if v_t represents the capsule output at time t, the BiLSTM learns temporal features as:

$$\overrightarrow{h_t} = LSTM_f\left(v_t, \overrightarrow{h_{t-1}}\right), \tag{13}$$

$$\overleftarrow{h_t} = LSTM_b\left(v_t, \overrightarrow{h_{t+1}}\right), \tag{14}$$

$$\overleftrightarrow{h_t} = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]. \tag{15}$$

This bidirectional formulation allows the model to analyze both past and future contextual dependencies, ensuring a holistic understanding of temporal behavior across frame sequences. The BiLSTM output is subsequently passed to the Global Average Pooling (GAP) layer for dimensionality reduction and further classification processing.

3.2.4. GAP Layer. The GAP layer replaces traditional fully connected layers by computing the average of each feature map in the final convolutional layer. For a feature map $F \in \mathbb{R}^{h \times w}$, GAP is computed as:

$$GAP(F) = \frac{1}{h.w} \sum_{i=1}^{h} \sum_{j=1}^{w} F_{ij}, \tag{16}$$

where F denotes the feature map, $h.w$ is the height and width of the feature map, and F_{ij} is the value at pixel (i, y). This significantly reduces the number of trainable parameters, thereby minimizing overfitting and preserving global contextual information.

3.2.5. Flatten Layer. Following GAP, the flatten layer reshapes the output into a one-dimensional vector. It transforms the tensor from shape $(batch_{size}, channels, 1{,}1)$ into $(batch_{size}, channels)$, serving as a bridge between convolutional feature extractors and subsequent high-level reasoning layers.

3.2.6. Fully Connected (Dense) Layer. The dense layer functions as a fully connected neural layer, responsible for learning complex feature representations. It performs a linear transformation followed by a non-linear activation:

$$y = f(Wx + b), \tag{17}$$

where x is the input vector, W is the weight matrix, b is the bias vector, and f is an activation function. This layer enhances feature abstraction and captures semantic information crucial for distinguishing deepfakes.

3.2.7. Sigmoid Activation Layer. The final layer employs the sigmoid activation function $\sigma(x)$ for binary classification:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{18}$$

This produces a probability score $p \in [0,1]$, where: $p \approx 1 \rightarrow real$, $p \approx 0 \rightarrow likely\ fake$.

It is compatible with binary cross-entropy loss, making it ideal for binary classification problems.

Table 2. Input-Output Tensor Flow and Integration with MediaPipe

Stage	Input Tensor	Output Tensor
BlazeFace	128 × 128 × 3	[N, 4] bounding boxes + [N, 6, 2] keypoints
MediaPipe Face Mesh	128 × 128 × 3 (face crop)	[468, 3] landmarks
Xception	128 × 128 × 3	[batch_size, h, w, c] feature maps

3.2.8. Fusion of MediaPipe Landmarks with XceptionCapsule Pipeline. The extracted 3D facial landmarks from MediaPipe Face Mesh are used as auxiliary behavioral and geometric features. Each frame's landmark vector $L = [x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_{468}, y_{468}, z_{468}] \in \mathbb{R}^{1404}$ is normalized and concatenated with the high-level spatial features $F_s \in \mathbb{R}^{h \times w \times c}$ obtained from the Xception encoder.

To ensure compatibility in feature space, a linear projection layer $\phi(\cdot)$ maps landmark features into the same latent dimension:

$$F_l = \emptyset(L) = W_l L + b_l, \tag{19}$$

where $W_l \in \mathbb{R}^{d \times 1404}$ and $b_l \in \mathbb{R}^d$. The projected landmark vector is then fused with the Xception feature tensor using an additive attention-based blending:

$$F_{fusion} = \propto . F_s + (1 - \propto) . F_l, \tag{20}$$

where $\propto \in [0,1]$ controls the contribution of spatial and landmark features, optimized during training. The fused representation F_{fusion} is passed to the CapsNet module, enabling the model to jointly reason about spatial, geometric, and behavioral cues. This process enhances sensitivity to micro-

expressions, eye-blink patterns, and lip-sync deviations that are often missed by frame-only CNNs.

4. Results and Discussion. This section presents the evaluation of the proposed deep learning model for deepfake detection in facial videos. By leveraging spatial and temporal features in facial expressions and movements, the model accurately distinguishes authentic from manipulated content. The approach achieves high detection performance while maintaining interpretability and reliability – essential for real-world applications such as media forensics, digital trust frameworks, and security systems.

4.1. Experimental Setup and Configuration. The proposed deepfake detection framework was implemented in Python 3.10 using Jupyter Notebook. TensorFlow was used for model development and training, while Scikit-learn handled data preprocessing, analysis, and evaluation. Experiments were conducted on a 64-bit Windows 10 system with a 7th-generation Intel Core i7 (2.8 GHz). Datasets were split 80/20 for training and testing, providing a stable platform for evaluating complex architectures such as XceptionCapsule Net.

4.2. Dataset Description.

Dataset 1: FaceForensics++ (FF++): The FF++ dataset [33] serves as a comprehensive and widely adopted benchmark for deepfake detection research. It comprises approximately 1,000 real video samples sourced from YouTube, each containing clear, stable, and front-facing facial recordings suitable for manipulation. From these original videos, over 4,000 manipulated videos were synthesized using four distinct facial manipulation techniques: DeepFakes (autoencoder-based face replacement), FaceSwap (graphics-based replacement), Face2Face (real-time facial reenactment), and NeuralTextures (neural rendering-based synthesis). A notable strength of FF++ is its provision of multiple compression settings, including raw (uncompressed), high-quality (HQ), and low-quality (LQ). This simulates real-world video scenarios, such as those encountered on social media platforms. Additionally, FF++ includes bounding box annotations and, in some versions, manipulated region masks, thereby supporting extended tasks such as forgery localization in addition to classification. With separate training, validation, and test splits, FF++ provides a standardized and reproducible benchmark for evaluating deepfake detection models under varying levels of manipulation complexity and video quality.

Dataset 2: Celeb-DF(v2). The Celeb-DF (v2) dataset [34], short for Celebrity DeepFake Dataset, addresses the limitations found in earlier datasets by offering more realistic and visually coherent deepfake videos. It consists of approximately 590 authentic video clips featuring 59 public figures, sourced from YouTube interviews that capture a diverse range of

head poses, facial expressions, and lighting conditions. A total of 5,639 deepfake videos were produced from these real videos using advanced deepfake synthesis methods, which significantly reduce visual artifacts such as lip mismatches, flickering, and facial warping. Unlike FF++, Celeb-DF (v2) focuses solely on high-quality deepfakes without introducing artificial compression, making it more representative of real-world high-fidelity forgeries. Although it lacks mask or manipulation annotations, its high visual fidelity and absence of synthetic glitches make it a valuable dataset for testing the generalization capability and robustness of detection models trained on noisier or lower-quality datasets. Together, FF++ and Celeb-DF (v2) provide complementary testbeds, one offering variety in manipulation and compression, and the other emphasizing realism and subtlety, making them suitable for the comprehensive evaluation of deepfake detection frameworks.

4.3. Hyperparameter Configuration. The training process was carefully configured using a standard set of hyperparameters to optimize model performance while mitigating overfitting and ensuring generalizability. The key hyperparameter settings used in the proposed framework are summarized in Table 2.

Table 2. Hyperparameter configuration

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	30
Learning rate Scheduler	ReduceLROnPlateau
Regularization	L2
Dropout	0.5
Validation split	5-fold

The Adam optimizer [35] was selected for its ability to adapt the learning rate during training. A learning rate scheduler was employed to adjust the rate dynamically in response to changes in validation loss. L2 regularization and dropout [36] were applied to reduce overfitting. A 5-fold cross-validation strategy was used to ensure robust performance estimation and evaluate the model's generalization across different subsets of the data.

4.4. Results and Analysis of the Preprocessing Stage. The preprocessing stage is crucial for ensuring accurate and reliable deepfake detection. Input videos are first decomposed into frames at uniform intervals

to preserve temporal coherence. BlazeFace detects faces under varying conditions, and MediaPipe Face Mesh extracts detailed facial landmarks for precise alignment. This ensures high-quality, geometrically consistent facial regions are passed to the XceptionCapsule Net, allowing the model to focus on subtle manipulations such as unnatural warping, inconsistent expressions, and temporal irregularities. Figures 5–8 illustrate successful face detection and alignment on the FF++ and Celeb-DF (v2) datasets, demonstrating the effectiveness of the preprocessing pipeline.

Fig. 5. Preprocessing result of BlazeFace and FaceMesh of fake videos for FF++ dataset

Fig. 6. Preprocessing result of BlazeFace and FaceMesh of real videos for FF++ dataset

Fig. 7. Preprocessing result of BlazeFace and FaceMesh of fake videos for Celeb-DF (v2) dataset

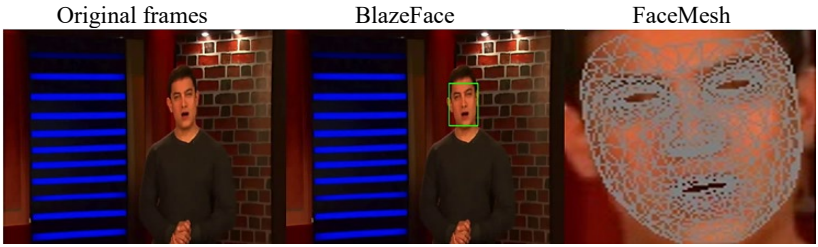


Fig. 8. Preprocessing result of BlazeFace and FaceMesh of real videos for Celeb-DF (v2) dataset

4.5. Performance Metrics. Performance metrics play a crucial role in assessing the effectiveness of deepfake detection models. A comprehensive set of metrics is used to quantify how well the model differentiates between authentic and manipulated content. These include classification metrics such as Accuracy, Precision, Recall, F1-Score, and Specificity. Collectively, these metrics provide detailed insights into the model's prediction quality, robustness, and reliability under real-world conditions.

The metrics used are defined as follows:

i. Accuracy. The proportion of correctly classified samples (real or fake) among all samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{21}$$

ii. Precision. The proportion of samples predicted as fake that are truly fake:

$$Precision = \frac{TP}{TP + FP}. \tag{22}$$

iii. Recall. The proportion of actual fake samples that are correctly identified:

$$Recall = \frac{TP}{TP + FN}. \tag{23}$$

iv. F1-Score. The harmonic mean of precision and recall, useful for imbalanced datasets:

$$F1 - Score = \frac{2.(Precision.Recall)}{Precision + Recall}. \quad (24)$$

v. Specificity. The proportion of real (negative class) samples correctly identified:

$$Specificity = \frac{TN}{TN + FP}. \quad (25)$$

vi. Area under the ROC Curve (AUC). This metric is typically used with the ROC (Receiver Operating Characteristic) curve to evaluate binary classifiers. It measures the model's ability to distinguish between the two classes (e.g., real vs. fake). Mathematically, it can be expressed as follows:

$$AUC = \int_0^1 TPR(FPR)d(FPR), \quad (26)$$

where TPR and FPR are the True Positive Rate and False Positive Rate at threshold i, and n is the number of thresholds used.

4.5.1. Cross-Validation Evaluation. To ensure robustness and generalization, a 5-fold cross-validation strategy is adopted. The dataset is partitioned into five equally sized subsets (folds). In each iteration, four folds are used for training, and the remaining fold is used for testing. This process is repeated five times, ensuring each fold serves exactly once as a test set. The final performance metrics are computed as the average of results across all folds, thereby reducing both variance and bias in model evaluation. This cross-validation approach is particularly valuable for deepfake detection, where data heterogeneity (in lighting, pose, and manipulation type) can influence model behavior. Table 3 presents the detailed results obtained from the 5-fold cross-validation on the FF++ dataset and Celeb-DF (v2) datasets, respectively. Additionally, Figure 9 illustrates the training and validation accuracy and loss curves across the five folds for the FF++ dataset, providing insights into model convergence and overfitting tendencies.

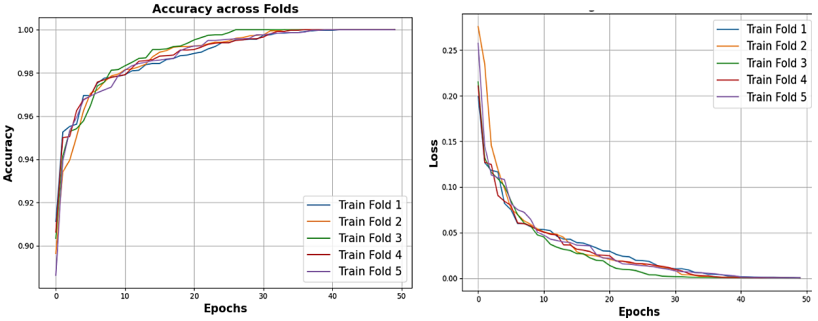


Fig. 9. Cross-validation for the accuracy and loss for FF++ dataset

Table 2. Cross-Validation Performance of the Proposed Model on FF++ and Celeb-DF (v2) Datasets

Datasets	Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Mean ± 95% CI
FF++	Accuracy	0.9667	0.9771	0.9510	0.9708	0.9708	0.9673	0.9673 ± 0.009
	Precision	0.9737	0.9718	0.9456	0.9733	0.9753	0.9679	0.9679 ± 0.011
	Recall	0.9612	0.9837	0.9591	0.9693	0.9673	0.9681	0.9681 ± 0.009
	F1-score	0.9671	0.9777	0.9523	0.9713	0.9713	0.9679	0.9679 ± 0.010
	Specificity	0.9724	0.9703	0.9427	0.9724	0.9745	0.9665	0.9665 ± 0.012
	AUC	0.9978	0.9962	0.9938	0.9957	0.9970	0.9961	0.9961 ± 0.001
Celeb-DF v2	Accuracy	0.9902	0.9872	0.9931	0.9921	0.9843	0.9894	0.9894 ± 0.003
	Precision	0.9987	0.9778	0.9906	0.9943	0.9923	0.9907	0.9907 ± 0.008
	Recall	0.9924	0.9981	0.9962	0.9905	0.9773	0.9909	0.9909 ± 0.007
	F1-score	0.9905	0.9878	0.9934	0.9924	0.9848	0.9898	0.9898 ± 0.005
	Specificity	0.9877	0.9754	0.9897	0.9938	0.9918	0.9877	0.9877 ± 0.007
	AUC	0.9998	0.9992	0.9999	0.9992	0.9989	0.9994	0.9994 ± 0.001

Table 2 presents the 5-fold cross-validation results of the proposed deepfake detection model on the FF++ and Celeb-DF (v2) datasets, including average values and 95% confidence intervals. For FF++, the model achieved an average accuracy of 96.73% with a mean AUC of 0.9961, demonstrating strong discriminative power. Precision, recall, and F1-score were consistently high across all folds, with mean values of 96.79%, 96.81%, and 96.79%, respectively. Similarly, on the Celeb-DF (v2) dataset, the model achieved a mean accuracy of 98.94% and an AUC of 0.9994, with precision, recall, and F1-score all above 98.9%, indicating highly reliable performance. The low 95% confidence intervals across all metrics highlight the model's robustness and consistent performance across different folds, confirming its generalization capability across diverse datasets and challenging deepfake scenarios. Figure 10 represents the cross-validation training and validation accuracy and loss for Celeb-DF (v2)

dataset, further corroborating the model's stable performance across different datasets and data distributions.

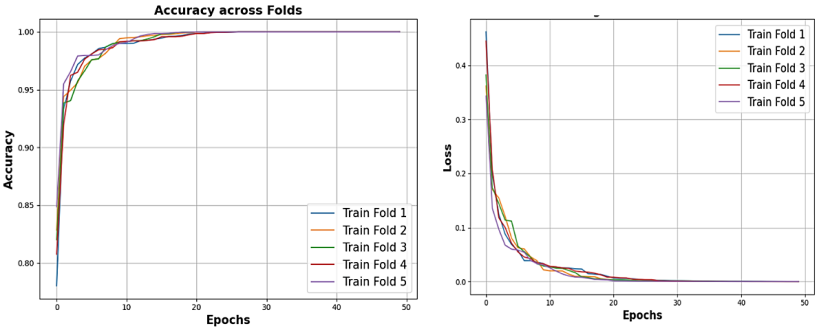


Fig. 10. Cross-validation for the accuracy and loss for Celeb-DF (v2) dataset

Figures 11 and 12 illustrate the training and validation accuracy and loss curves for the FF++ and Celeb-DF (v2) datasets, respectively, providing visual confirmation of the model's convergence and generalization effectiveness across both datasets.

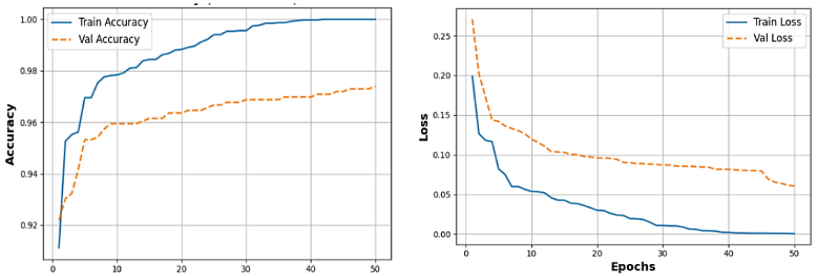


Fig. 11. Training and validation accuracy and loss of dataset 1

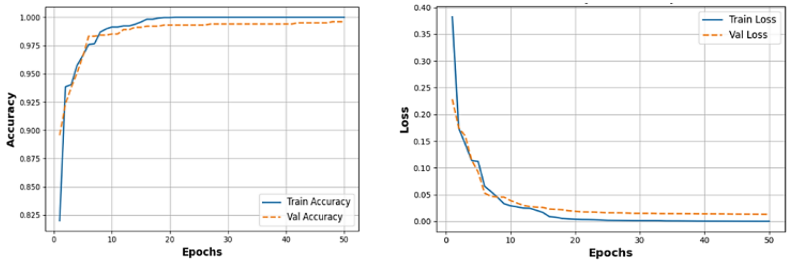
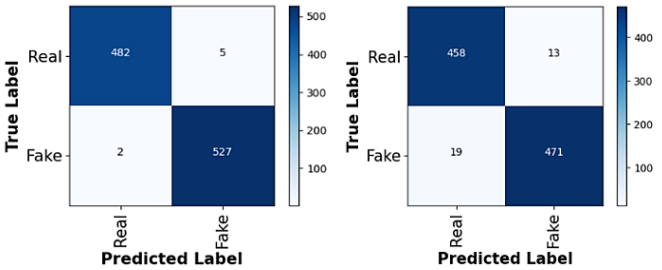


Fig. 12. Training and validation accuracy and loss of dataset 2

4.5.2. Confusion Matrix. A confusion matrix serves as a valuable tool for assessing the performance of classification models, providing an in-depth comparison between actual and predicted class labels. It categorizes outcomes into True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), facilitating accurate computation of metrics such as accuracy, precision, recall, and error rates. This visualization is especially useful for evaluating how well a model differentiates between classes, making it highly effective in binary classification scenarios such as deepfake detection.

Figure 13 presents the confusion matrices of the Dataset 1 and the Dataset 2.



a) b)

Fig. 13. Confusion Matrix: a) dataset 1, b) dataset 2

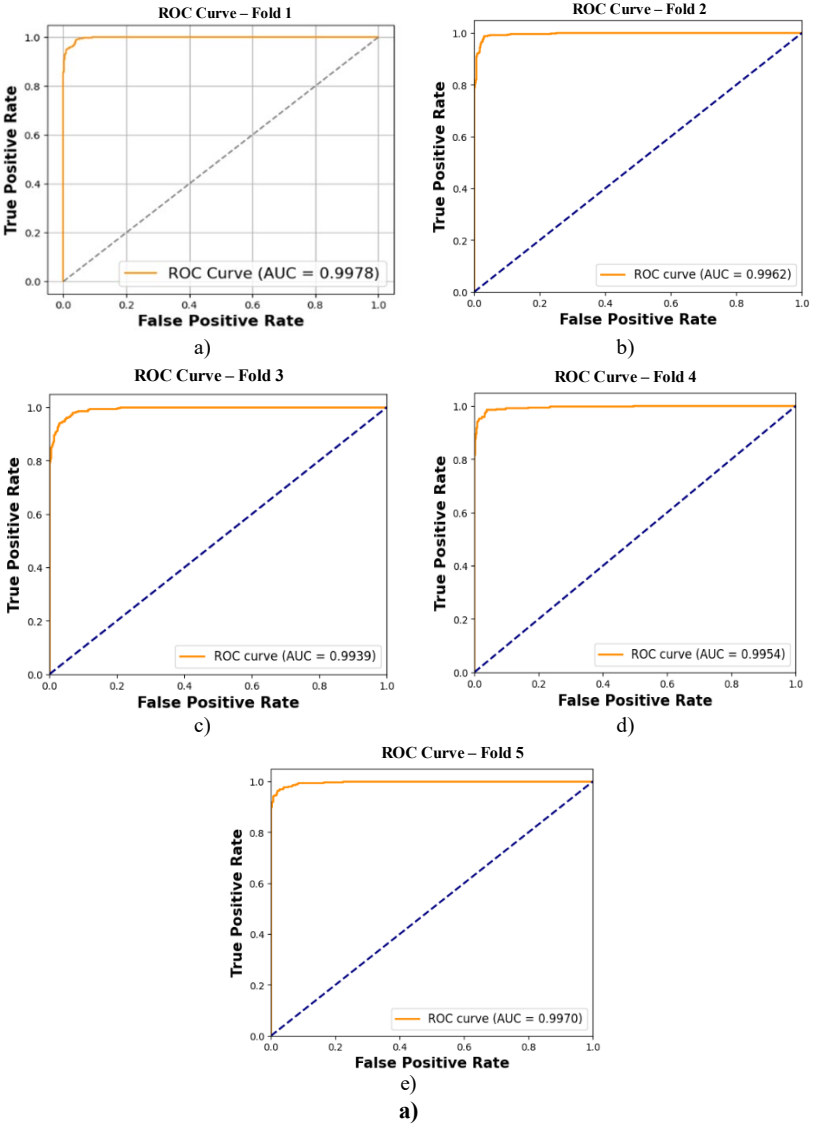
For Dataset 1(a), the model accurately classified 482 real and 527 fake instances. The model misclassified only 5 genuine instances as fake and incorrectly identified 2 manipulated samples as authentic. This distribution reflects excellent model performance with minimal misclassification, indicating high precision, recall, and overall robustness in detecting facial manipulations.

For Dataset 2(b), the model correctly identified 458 real and 471 fake videos. However, it misclassified 13 real instances as fake and 19 fake instances as real, reflecting a slightly higher error rate compared to Dataset 1. Despite this, the matrix exhibits strong diagonal dominance, affirming the model's capacity to generalize well to more challenging and realistic deepfake samples.

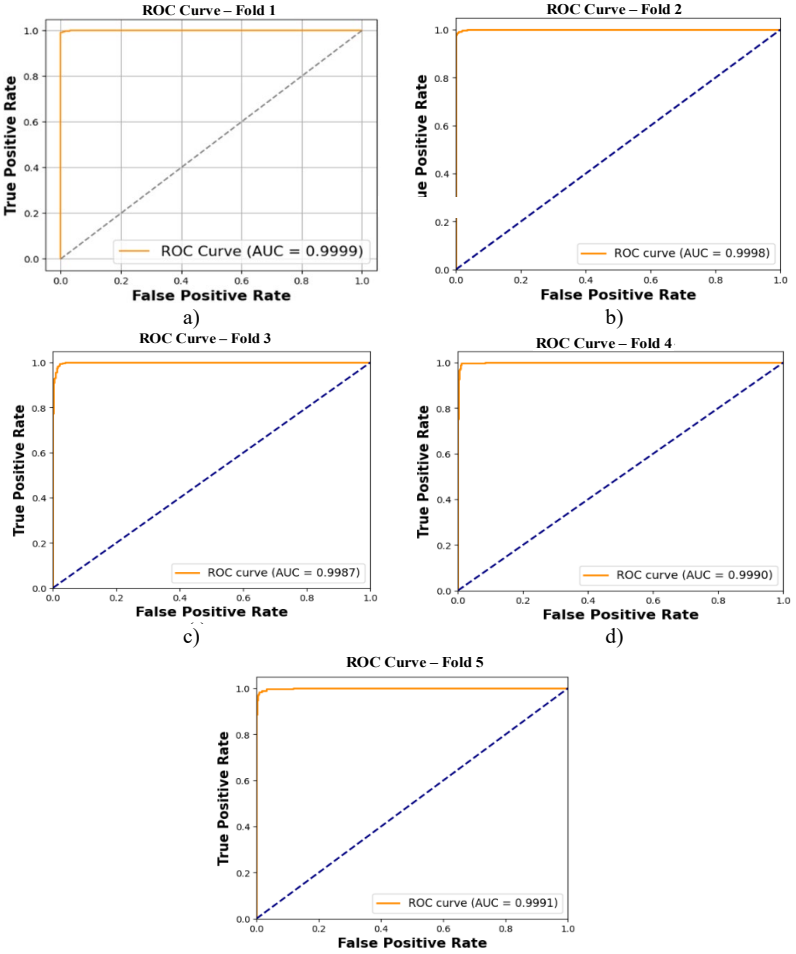
Overall, the confusion matrices underscore the model's high discriminative power and low error rate, particularly with Dataset 1, while still maintaining robust performance on the more visually complex Dataset 2.

4.5.3. ROC Curve. The ROC curve illustrates the relationship between the TPR, also known as sensitivity, and the FPR across varying classification thresholds. By visualizing this trade-off, the ROC curve provides insight into the model's ability to discriminate between classes regardless of the decision threshold. A key metric derived from the ROC curve is the AUC, which offers a

single scalar value to summarize performance. An AUC value closer to 1.0 indicates a high degree of separability between the positive and negative classes, signifying excellent classification capability. Figure 14 shows the ROC curves for (a) FF++ dataset and (b) Celeb-DF (v2) dataset.



a)



b)

Fig. 14. ROC curve for: a) FF++ dataset, b) Celeb-DF (v2) dataset

The ROC curves, averaged across the 5-fold cross-validation, demonstrate the model's exceptional classification ability. For the FF++ dataset (a), the model achieved an average AUC of 0.9961, indicating a very high true positive rate with a minimal false positive rate across all folds. For the Celeb-DF (v2) dataset (b), the model attained an average AUC of 0.9994, reflecting near-perfect separability between real and manipulated content. These ROC curves, based on averaged cross-validation results, reaffirm the

high discriminative power and robustness of the proposed framework across different datasets, diverse conditions, and various manipulation techniques.

4.6. Comparative Analysis. This section presents a comprehensive comparative analysis of the proposed deepfake detection framework against a range of existing state-of-the-art approaches. The evaluation considers two primary performance metrics such as Accuracy and AUC measured across two benchmark datasets: FF++ and Celeb-DF (v2). The compared methods include both conventional CNN models (e.g., VGG16, InceptionV3, ResNet50, MobileNetV2, EfficientNet variants) and advanced architectures that employ hybrid learning, attention, and frequency-domain features such as MesoNet, F3-Net, RFM, GRAM, GFFD, SPSL, M2TR, GocNet, F2-Trans, BMNet, and Self-Supervised BEiT-HPR. Tables 5 and 6 summarize the comparative results on both datasets.

Table 5. The overall comparison for the FF++ dataset

Methods	FF++ dataset	
	Accuracy (%)	AUC(%)
Face cutout [23]	84	80
DWT [24]	95.13	95.49
ResNet-Swish-BiLSTM [25]	96.23	-
BMNet [26]	95.54	98.60
MLP [27]	98.00	99.94
Self-Supervised BEiT-HPR [28]	83.92	-
ConvNext-PNe [29]	97.78	99.25
CViT2 [30]	94.80	96.00
MesoNet [31]	60.51	74.55
MesoNet-Inc4 [31]	82.15	83.64
Xception [31]	89.84	98.14
EfficientNetb4 [31]	91.89	98.45
F3-Net [31]	93.78	98.55
RFM [31]	91.59	98.37
GRAM [31]	92.21	97.81
GFFD [31]	90.23	98.28
SPSL [31]	91.50	95.32
M2TR [31]	94.08	98.43
GocNet [31]	91.67	97.58
F2-Trans [31]	96.60	99.24
MSTN [31]	95.78	95.78
VGG16 [32]	78.39	78.00
InceptionV3 [32]	51.00	50.00
ResNet50 [32]	89.67	89.00
Xception [32]	73.86	73.00
MobileNetV2 [32]	76.63	76.00
EfficientNetB7 [32]	83.66	84.00
Proposed	96.67	99.78

Table 6. The overall comparison for the Celeb-DF (v2) dataset

Methods	Celeb-DF (v2) dataset	
	Accuracy (%)	AUC(%)
Face cutout [23]	92	93
DWT [24]	95.49	95.49
ResNet-Swish-BiLSTM [25]	78.33	-
BMNet [26]	80.20	75.72
MLP [27]	96.33	97.05
Self-Supervised BEiT-HPR [28]	98.25	-
ConvNext-PNe [29]	97.09	98.99
CViT2 [30]	98.30	99.00
FWA [31]	50.74	55.71
CviT [31]	51.67	60.22
Capsule [31]	63.35	62.70
MesoNet [31]	42.80	52.31
Xception [31]	55.92	67.23
BMNet [31]	80.20	75.72
VGG16 [32]	67.09	68.00
InceptionV3 [32]	55.12	52.00
ResNet50 [32]	67.09	68.00
Xception [32]	59.22	63.00
MobileNetV2 [32]	63.24	65.00
EfficientNetB7 [32]	70.08	69.00
Proposed	99.31	99.99

Footnote: The accuracy and AUC values for the proposed method correspond to the best-performing fold (Full Model) from the cross-validation experiment, used for fair comparison with state-of-the-art single-model results. Mean accuracy across all five folds (98.94%) is reported separately in Table 2, confirming consistent model performance and stable generalization.

Table 5 illustrates the comparative performance of different deepfake detection methods on the FF++ dataset. The proposed model achieves the highest performance, with an accuracy of 96.67% and an AUC of 99.78%, thereby outperforming both classical CNN-based methods and modern hybrid architectures. Traditional CNN-based models such as VGG16, InceptionV3, and MobileNetV2 show moderate performance with accuracies between 51% and 83%, indicating their limited ability to capture complex spatial-temporal inconsistencies present in manipulated videos. Recent architectures such as MesoNet, MesoNet-Inc4, and Xception deliver improved results but still fall short in generalization across diverse forgery

types. Advanced techniques such as EfficientNetB4, F3-Net, RFM, GRAM, M2TR, and F2-Trans achieve higher AUC values (above 98%), reflecting progress in feature extraction and classification precision. However, the proposed XceptionCapsule-based framework outperforms all competitors, exhibiting near-perfect discriminative capability as evidenced by its 99.78% AUC. This superior performance can be attributed to the model's ability to jointly capture spatial texture cues and temporal motion inconsistencies, while the integrated Capsule Network layer enhances dynamic feature representation and resilience to occlusions and compression artifacts.

Table 6 presents the comparative evaluation of the proposed model against state-of-the-art methods on the Celeb-DF (v2) dataset, which is known for its high-quality and challenging deepfake content. The proposed framework demonstrates remarkable superiority, achieving an accuracy of 99.31% and an AUC of 99.99%, substantially surpassing all prior methods. While traditional CNN-based detectors such as ResNet50, VGG16, and InceptionV3 achieve accuracies below 70%, specialized architectures such as BMNet, CViT, and Capsule Networks also fail to generalize effectively to complex manipulations, with most methods performing below 81% accuracy. On the other hand, transformer-based and hybrid learning approaches, such as ConvNext-PNet, CViT2, and Self-Supervised BEiT-HPR, show improved performance with accuracies exceeding 97%, yet they still fall short of the proposed model's near-perfect results. The performance of the proposed model on Celeb-DF (v2) confirms its robust generalization ability and strong resistance to unseen manipulations, even under high-fidelity synthesis and compression variations.

4.7. Ablation Study. To assess the contribution of each module within the proposed framework, an ablation experiment was conducted on the FF++ and Celeb-DF (v2) datasets. The analysis evaluates performance gains achieved by progressively integrating CapsNet, BlazeFace, MediaPipe FaceMesh, and the BiLSTM layer with the baseline Xception model.

Table 7 presents the ablation study results illustrating the contribution of each component integrated into the proposed deepfake detection framework. The baseline Xception model achieved 94.26% accuracy and 97.85% AUC, establishing the initial performance. Incorporating the CapsNet module improved feature representation and slightly increased overall accuracy to 96.03%, with an AUC of 98.91%. When combined with BlazeFace, the framework achieved improved face localization and yielded 96.30% accuracy and 99.23% AUC. The integration of MediaPipe FaceMesh further enhanced landmark precision, resulting in 96.50% accuracy and 99.64% AUC. When the BiLSTM layer was introduced, the model achieved 96.58% accuracy and 99.70% AUC,

confirming its contribution in learning temporal dependencies and motion-based cues such as blinking, head rotation, and lip synchronization. Finally, the full model, integrating all modules (Xception, CapsNet, BlazeFace, MediaPipe Face Mesh, and BiLSTM), attained the highest performance: 96.67% accuracy, 97.31% precision, 96.12% recall, 96.71% F1-score, and 99.78% AU. This demonstrates that the synergistic integration of spatial, geometric, and temporal components significantly enhances deepfake detection reliability and robustness.

Table 7. Ablation Study Showing the Impact of Different Model Components on Deepfake Detection Performance on FF++ Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Xception only	94.26	93.40	95.10	94.24	97.85
Xception + CapsNet	96.03	95.71	95.80	96.07	98.91
XceptionCapsule + BlazeFace only	96.30	96.95	95.95	96.20	99.23
XceptionCapsule + MediaPipe FaceMesh	96.50	97.20	96.05	96.40	99.64
XceptionCapsule + BiLSTM	96.58	97.08	96.11	96.45	99.70
Full Model	96.67	97.31	96.12	96.71	99.78

On the Celeb-DF (v2) dataset (Table 2), which consists of high-fidelity and visually coherent deepfakes, the incremental impact of each module is even more pronounced. The baseline Xception model achieved 92.45% accuracy and 96.72% AUC, struggling to capture subtle manipulations. The addition of CapsNet improved the detection of fine-grained artifacts, reaching 95.84% accuracy. Incorporating BlazeFace enhanced face localization accuracy, achieving an accuracy of over 97%. MediaPipe Face Mesh further refined geometric consistency, leading to 98.72% accuracy and 99.73% AUC. Introducing the BiLSTM module significantly improved temporal understanding and motion-based detection, achieving 99.01% accuracy and 99.88% AUC by modeling frame-to-frame dependencies and identifying inconsistencies in facial dynamics. The full model achieved the best results, with 99.31% accuracy and nearly perfect AUC (99.99%), confirming that combining hierarchical spatial modeling, geometric precision, and temporal learning substantially improves deepfake detection performance and generalization in complex real-world scenarios.

Table 8. Ablation Study Showing the Impact of Different Model Components on Deepfake Detection Performance on Celeb-DF (v2) Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Xception only	92.45	91.62	93.01	92.30	96.72
Xception + CapsNet	95.84	95.41	95.97	95.69	98.44
XceptionCapsule + BlazeFace only	97.25	97.10	97.42	97.26	99.11
XceptionCapsule + MediaPipe FaceMesh	98.72	98.45	98.80	98.62	99.73
XceptionCapsule + BiLSTM	99.01	98.80	99.25	99.02	99.88
Full Model	99.31	99.06	99.62	99.34	99.99

4.8. Discussion. The proposed Face and Motion-Aware Detection Framework effectively integrates spatial and temporal features to detect subtle facial artifacts and motion inconsistencies in deepfakes. BlazeFace provides robust face localization under varying conditions, and MediaPipe Face Mesh ensures precise landmark extraction. The XceptionCapsule Net preserves hierarchical spatial relationships, enabling the detection of fine-grained forgery cues. Additionally, the integration of a BiLSTM layer enables the framework to model temporal dependencies by learning motion dynamics such as blinking, lip movement, and head rotation across consecutive frames, enhancing temporal consistency in detection. The experimental results on the FF++ and Celeb-DF (v2) datasets demonstrate strong generalization, outperforming several state-of-the-art methods. Ablation studies confirm the contribution of each module, highlighting the robustness and adaptability of the framework to different manipulation techniques and challenging conditions such as low resolution, compression, and subtle manipulations. These characteristics position the framework as a promising solution for real-world media forensics and content authentication.

5. Conclusion. This study introduces a robust Face and Motion-Aware Detection Framework that advances deepfake detection by combining spatial and temporal analysis. Benchmark evaluations demonstrate its superior performance and generalization compared to existing methods. The framework's design allows adaptability to various types of manipulations and real-world conditions. For future work, the framework can be extended to multi-modal analysis by incorporating audio and textual cues, which would improve the detection of lip-sync or dialogue-incoherent deepfakes. Further enhancements could include transformer-based architectures to strengthen temporal modeling. Finally,

testing on in-the-wild deepfake samples and evaluating the feasibility of real-time deployment on social media or video streaming platforms are crucial steps for ensuring the practical applicability of the system against increasingly sophisticated generative manipulations.

References

1. O'Toole A.J., Castillo C.D. Face recognition by humans and machines: three fundamental advances from deep learning. Annual Review of Vision Science. 2021. vol. 7(1). pp. 543–570.
2. Fola-Rose A., Solomon E., Bryant K., Woubie A. A Systematic Review of Facial Recognition Methods: Advancements, Applications, and Ethical Dilemmas. IEEE International Conference on Information Reuse and Integration for Data Science (IRI). 2024. pp. 314–319.
3. EL Fadel N. Facial Recognition Algorithms: A Systematic Literature Review. Journal of Imaging. 2025. vol. 11(2).
4. Wu Y. Facial Recognition Technology: College Students' Perspectives in the US. Trends in Sociology. 2024. vol. 2(2). pp. 56–69. DOI: 10.61187/ts.v2i2.119.
5. Dang M., NguyenT.N. Digital face manipulation creation and detection: A systematic review. Electronics. 2023. vol. 12(16).
6. Masood M., Nawaz M., Malik K.M., Javed A., Irtaza A., Malik H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence. 2023. vol. 53(4). pp. 3974–4026.
7. Akkem Y., Biswas S.K., Varanasi A. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Engineering Applications of Artificial Intelligence. 2024. vol. 131.
8. Alqahtani H., Kavakli-Thorne M., Kumar G. Applications of generative adversarial networks (gans): An updated review. Archives of Computational Methods in Engineering. 2021. vol. 28. pp. 525–552.
9. Sun X., Chen S., Yao T., Liu H., Ding S., Ji R. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. Advances in Neural Information Processing Systems. 2024. vol. 37. pp. 101474–101497.
10. Ge Y., Xu J., Zhao B.N., Joshi N., Itti L., Vineet V. Dall-e for detection: Language-driven compositional image synthesis for object detection. arXiv preprint arXiv:2206.09592. 2022.
11. Zhao H., Liang T., Davari S., Kim D. Synthesizing Reality: Leveraging the Generative AI-Powered Platform Midjourney for Construction Worker Detection. arXiv preprint arXiv:2507.13221. 2025.
12. Zhou K.Z., Choudhry A., Gumusel E., Sanfilippo M.R. Sora is Incredible and Scary": Emerging Governance Challenges of Text-to-Video Generative AI Models. arXiv preprint arXiv:2406.11859. 2024.
13. Qadir A., Mahum R., El-Meligy M.A., Ragab A.E., AlSalman A., Awais M. An efficient deepfake video detection using robust deep learning. Heliyon, 2024. vol. 10(5).
14. Bhattacharyya C., Wang H., Zhang F., Kim S., Zhu X. Diffusion deepfake. arXiv preprint arXiv:2404.01579. 2024.
15. Al-Khazraji S.H., Saleh H.H., Khalid A.I., Mishkhal I.A. Impact of deepfake technology on social media: Detection, misinformation and societal implications. The Eurasia Proceedings of Science Technology Engineering and Mathematics. 2023. vol. 23. pp. 429–441.

16. Jbara W.A., Hussein N.A.H.K., Soud J.H. Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis. Mesopotamian Journal of CyberSecurity. 2024. vol. 4(3). pp. 233–250.
17. Arya M., Goyal U., Chawla S. A Study on Deep Fake Face Detection Techniques. 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2024. pp. 459–466.
18. Rajeev A., Raviraj P. An insightful analysis of digital forensics effects on networks and multimedia applications. SN Computer Science. 2023. vol. 4(2).
19. Sheremet O.I., Sadovoi O.V., Harshanov D.V., Kovalchuk S., Sheremet K.S., Sokhina Y.V. Efficient face detection and replacement in the creation of simple fake videos. Applied Aspects of Information Technology. 2023. vol. 6(3). pp. 286–303.
20. Wang J., Yuan S., Lu T., Zhao H., Zhao Y. Video-based real-time monitoring of engagement in E-learning using MediaPipe through multi-feature analysis. Expert Systems with Applications. 2025. vol. 288. DOI: 10.1016/j.eswa.2025.128239.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 1251–1258.
22. Joshi D., Kashyap A., Arora P. CapsNet-Based Deep Learning Approach for Robust Image Forgery Detection. 10th International Conference on Signal Processing and Communication (ICSC). IEEE, 2025. pp. 308–314.
23. Alanazi F., Ushaw G., Morgan G. Improving detection of deepfakes through facial region analysis in images. Electronics. 2024. vol. 13(1). DOI: 10.3390/electronics13010126.
24. Hasanaath A.A., Luqman H., Katib R., Anwar S. FSBI: Deepfake detection with frequency enhanced self-blended images. Image and Vision Computing. 2025. vol. 154.
25. Qadir A., Mahum R., El-Meligy M.A., Ragab A.E., AlSalman A., Awais M. An efficient deepfake video detection using robust deep learning. Heliyon, 2024. vol. 10(5).
26. Xiong D., Wen Z., Zhang C., Ren D., Li W. BMNet: Enhancing Deepfake Detection through BiLSTM and Multi-Head Self-Attention Mechanism. IEEE Access. 2025. vol. 13. pp. 21547–21556. DOI: 10.1109/ACCESS.2025.3533653.
27. Naskar G., Mohiuddin S., Malakar S., Cuevas E., Sarkar R. Deepfake detection using deep feature stacking and meta-learning. Heliyon. 2024. vol. 10(4).
28. Al Redhaei A., Fraihat S., Al-Betar M.A. A self-supervised BEiT model with a novel hierarchical patchReducer for efficient facial deepfake detection. Artificial Intelligence Review. 2025. vol. 58(9).
29. Ilyas H., Javed A., Malik K.M. ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection. IEEE International Joint Conference on Biometrics (IJCB). 2024. pp. 1–9.
30. Deressa D.W., Lambert P., Van Wallendael G., Atnafu S., Mareen H. Improved Deepfake Video Detection Using Convolutional Vision Transformer. IEEE Gaming, Entertainment, and Media Conference (GEM). 2024. pp. 1–6.
31. Gong R., He R., Zhang, D., Sangaiah A.K., Alenazi M.J. Robust face forgery detection integrating local texture and global texture information. EURASIP Journal on Information Security. 2025(1). vol. 3.
32. Saikia P., Dholaria D., Yadav P., Patel V., Roy M. A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. International joint conference on neural networks (IJCNN). IEEE, 2022. pp. 1–7.
33. Rossler A., Cozzolino D., Verdoliva L., Riess C., Thies J., et al., Faceforensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces. arXiv preprint arXiv:1803.09179. 2018.

34. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-df: A largescale challenging dataset for deepfake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 3207–3216.
35. Akbar A.F., Ayu P.D.W., Hostiadi D.P. Performance Analysis of Deep Learning Architectures in Classifying Fake and Real Images. JUITA: Jurnal Informatika. 2025. pp. 167–176.
36. Khan S.B., Gupta M., Gopinathan B., Thyluru RamaKrishna M., Saraee M., Mashat A., Almusharraf A. DeepFake Detection: Evaluating the Performance of EfficientNetV2-B2 on Real vs. Fake Image Classification. IET Image Processing. 2025. vol. 19(1).

Rajeev Aishwarya — Student, GSSS Institute of Engineering and Technology for Women. Research interests: deepfake detection, XceptionCapsule Net, Face Mesh, BlazeFace, facial landmark extraction, video forensics. The number of publications — 1. aishwaryarajeev654@gmail.com; KRS Road, Metagalli, 570016, Karnataka, Mysore, India; office phone: +91(0821)258-1305.

P. Raviraj — Professor, GSSS Institute of Engineering and Technology for Women. Research interests: deepfake detection, XceptionCapsule Net, Face Mesh, BlazeFace, facial landmark extraction, video forensics, image processing. The number of publications — 75. ravirajp654@gmail.com; KRS Road, Metagalli, 570016, Karnataka, Mysore, India; office phone: +91(0821)258-1305.

УДК 030 **DOI** 10.15622/ia.25.1.4

А. РАДЖИВ, РАВИРАДЖ П.

МОДЕЛЬ ДЛЯ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ С УЧЕТОМ ПРОСТРАНСТВЕННО-ВРЕМЕННЫХ И ПОВЕДЕНЧЕСКИХ ПРИЗНАКОВ НА ОСНОВЕ ОБЪЕДИНЕНИЯ XCEPTIONCAPSULE

Раджив А., Равирадж П. **Модель для обнаружения дипфейков с учетом пространственно-временных и поведенческих признаков на основе объединения XceptionCapsule.**

Аннотация. Обнаружение дипфейков по-прежнему представляет собой серьезную проблему, главным образом из-за ключевых ограничений существующих методов, включая зависимость от анализа отдельных кадров, уязвимость к видео низкого разрешения или сжатым видео, а также неспособность улавливать временные несоответствия. Кроме того, традиционные методы обнаружения лиц часто дают сбой в сложных условиях, таких как плохое освещение или окклюзия, а многие модели не справляются с тонкими манипуляциями из-за неадекватного извлечения признаков и переобучения на ограниченных наборах данных. Для устранения недостатков существующих подходов к обнаружению дипфейков в данном исследовании предлагается система обнаружения лиц и движений, которая объединяет как пространственную, так и временную информацию. Работа системы начинается с этапа предварительной обработки, на котором видеокадры извлекаются с фиксированной частотой для обеспечения временной согласованности. Области лица и детальные ориентиры точно определяются с помощью BlazeFace и MediaPipe Face Mesh. Затем эти признаки обрабатываются с помощью предлагаемой сети XceptionCapsule Net, которая сочетает в себе возможности извлечения пространственных признаков модели Xception с иерархическим и учитывающим ракурс представлением капсульных сетей (CapsNet), а также возможностью моделирования временных зависимостей двунаправленного слоя долгой краткосрочной памяти (BiLSTM). Архитектура включает в себя глобальный усредняющий пулинг, сглаживание и полносвязные слои с сигмоидной функцией активации для бинарной классификации. Обширные оценки на наборах данных FaceForensics++ (FF++) и Celeb-DF демонстрируют высокую производительность, достигая точности до 99,31% и площади под кривой (AUC) 99,99%. Результаты подтверждают эффективность, точность и обобщающую способность системы для видео различного качества и сценариев манипуляций.

Ключевые слова: обнаружение дипфейков, XceptionCapsule Net, Face Mesh, BlazeFace, извлечение лицевых ориентиров, видеокриминалистика.

Литература

1. O'Toole A.J., Castillo C.D. Face recognition by humans and machines: three fundamental advances from deep learning. Annual Review of Vision Science. 2021. vol. 7(1). pp. 543–570.
2. Fola-Rose A., Solomon E., Bryant K., Woubie A. A Systematic Review of Facial Recognition Methods: Advancements, Applications, and Ethical Dilemmas. IEEE International Conference on Information Reuse and Integration for Data Science (IRI). 2024. pp. 314–319.
3. EL Fadel N. Facial Recognition Algorithms: A Systematic Literature Review. Journal of Imaging. 2025. vol. 11(2).

4. Wu Y. Facial Recognition Technology: College Students' Perspectives in the US. Trends in Sociology. 2024. vol. 2(2). pp. 56–69. DOI: 10.61187/ts.v2i2.119.
5. Dang M., NguyenT.N. Digital face manipulation creation and detection: A systematic review. Electronics. 2023. vol. 12(16).
6. Masood M., Nawaz M., Malik K.M., Javed A., Irtaza A., Malik H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence. 2023. vol. 53(4). pp. 3974–4026.
7. Akkem Y., Biswas S.K., Varanasi A. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Engineering Applications of Artificial Intelligence. 2024. vol. 131.
8. Alqahtani H., Kavakli-Thorne M., Kumar G. Applications of generative adversarial networks (gans): An updated review. Archives of Computational Methods in Engineering. 2021. vol. 28. pp. 525–552.
9. Sun X., Chen S., Yao T., Liu H., Ding S., Ji R. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. Advances in Neural Information Processing Systems. 2024. vol. 37. pp. 101474–101497.
10. Ge Y., Xu J., Zhao B.N., Joshi N., Itti L., Vineet V. Dall-e for detection: Language-driven compositional image synthesis for object detection. arXiv preprint arXiv:2206.09592. 2022.
11. Zhao H., Liang T., Davari S., Kim D. Synthesizing Reality: Leveraging the Generative AI-Powered Platform Midjourney for Construction Worker Detection. arXiv preprint arXiv:2507.13221. 2025.
12. Zhou K.Z., Choudhry A., Gumusel E., Sanfilippo M.R. Sora is Incredible and Scary": Emerging Governance Challenges of Text-to-Video Generative AI Models. arXiv preprint arXiv:2406.11859. 2024.
13. Qadir A., Mahum R., El-Meligy M.A., Ragab A.E., AlSalman A., Awais M. An efficient deepfake video detection using robust deep learning. Heliyon, 2024. vol. 10(5).
14. Bhattacharyya C., Wang H., Zhang F., Kim S., Zhu X. Diffusion deepfake. arXiv preprint arXiv:2404.01579. 2024.
15. Al-Khazraji S.H., Saleh H.H., Khalid A.I., Mishkhal I.A. Impact of deepfake technology on social media: Detection, misinformation and societal implications. The Eurasia Proceedings of Science Technology Engineering and Mathematics. 2023. vol. 23. pp. 429–441.
16. Jbara W.A., Hussein N.A.H.K., Soud J.H. Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis. Mesopotamian Journal of CyberSecurity. 2024. vol. 4(3). pp. 233–250.
17. Arya M., Goyal U., Chawla S. A Study on Deep Fake Face Detection Techniques. 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2024. pp. 459–466.
18. Rajeev A., Raviraj P. An insightful analysis of digital forensics effects on networks and multimedia applications. SN Computer Science. 2023. vol. 4(2).
19. Sheremet O.I., Sadovoi O.V., Harshanov D.V., Kovalchuk S., Sheremet K.S., Sokhina Y.V. Efficient face detection and replacement in the creation of simple fake videos. Applied Aspects of Information Technology. 2023. vol. 6(3). pp. 286–303.
20. Wang J., Yuan S., Lu T., Zhao H., Zhao Y. Video-based real-time monitoring of engagement in E-learning using MediaPipe through multi-feature analysis. Expert Systems with Applications. 2025. vol. 288. DOI: 10.1016/j.eswa.2025.128239.
21. Chollet F. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 1251–1258.

22. Joshi D., Kashyap A., Arora P. CapsNet-Based Deep Learning Approach for Robust Image Forgery Detection. 10th International Conference on Signal Processing and Communication (ICSC). IEEE, 2025. pp. 308–314.
23. Alanazi F., Ushaw G., Morgan G. Improving detection of deepfakes through facial region analysis in images. Electronics. 2024. vol. 13(1). DOI: 10.3390/electronics13010126.
24. Hasanaath A.A., Luqman H., Katib R., Anwar S. FSBI: Deepfake detection with frequency enhanced self-blended images. Image and Vision Computing. 2025. vol. 154.
25. Qadir A., Mahum R., El-Meligy M.A., Ragab A.E., AlSalman A., Awais M. An efficient deepfake video detection using robust deep learning. Heliyon, 2024. vol. 10(5).
26. Xiong D., Wen Z., Zhang C., Ren D., Li W. BMNet: Enhancing Deepfake Detection through BiLSTM and Multi-Head Self-Attention Mechanism. IEEE Access. 2025. vol. 13. pp. 21547–21556. DOI: 10.1109/ACCESS.2025.3533653.
27. Naskar G., Mohiuddin S., Malakar S., Cuevas E., Sarkar R. Deepfake detection using deep feature stacking and meta-learning. Heliyon. 2024. vol. 10(4).
28. Al Redhaei A., Fraihat S., Al-Betar M.A. A self-supervised BEiT model with a novel hierarchical patchReducer for efficient facial deepfake detection. Artificial Intelligence Review. 2025. vol. 58(9).
29. Ilyas H., Javed A., Malik K.M. ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection. IEEE International Joint Conference on Biometrics (IJCB). 2024. pp. 1–9.
30. Deressa D.W., Lambert P., Van Wallendael G., Atnafu S., Mareen H. Improved Deepfake Video Detection Using Convolutional Vision Transformer. IEEE Gaming, Entertainment, and Media Conference (GEM). 2024. pp. 1–6.
31. Gong R., He R., Zhang, D., Sangaiah A.K., Alenazi M.J. Robust face forgery detection integrating local texture and global texture information. EURASIP Journal on Information Security. 2025(1). vol. 3.
32. Saikia P., Dholaria D., Yadav P., Patel V., Roy M. A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. International joint conference on neural networks (IJCNN). IEEE, 2022. pp. 1–7.
33. Rossler A., Cozzolino D., Verdoliva L., Riess C., Thies J., et al., Faceforensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces. arXiv preprint arXiv:1803.09179. 2018.
34. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-df: A largescale challenging dataset for deepfake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 3207–3216.
35. Akbar A.F., Ayu P.D.W., Hostiadi D.P. Performance Analysis of Deep Learning Architectures in Classifying Fake and Real Images. JUITA: Jurnal Informatika. 2025. pp. 167–176.
36. Khan S.B., Gupta M., Gopinathan B., Thyluru RamaKrishna M., Saraee M., Mashat A., Almusharraf A. DeepFake Detection: Evaluating the Performance of EfficientNetV2-B2 on Real vs. Fake Image Classification. IET Image Processing. 2025. vol. 19(1).

Раджив Айшвария — студент, Инженерный институт для женщин в Майсуре. Область научных интересов: обнаружение дипфейков, XceptionCapsule Net, Face Mesh, BlazeFace, извлечение лицевых ориентиров, видеокриминалистика. Число научных публикаций — 1. aishwaryarajeev654@gmail.com; KRS Роад, Метагалли, 570016, Карнатака, Майсур, Индия; р.т.: +91(0821)258-1305.

П. Равирадж — профессор, Инженерный институт для женщин в Майсуре. Область научных интересов: обнаружение дипфейков, XceptionCapsule Net, Face Mesh, BlazeFace, извлечение лицевых ориентиров, видеокриминалистика, обработка изображений. Число научных публикаций — 75. ravirajp654@gmail.com; KRS Роад, Метагалли, 570016, Карнатака, Майсур, Индия; р.т.: +91(0821)258-1305.