Journal of Theoretical and Applied Information Technology
15th April 2025. Vol.103. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EXPENDABLE DEEPFAKE DETECTION USING MULTI-STAGE DEEP LEARNING WITH SPATIAL, TEMPORAL, AND FREQUENCY FORENSIC PIPELINES

[1]SATHYAVANI ADDANKI,MANEESHA VADDURI, [2]VIJAYKIRAN ADDANKI,[3]NUTHALAPATI KAMALA VIKASINI

[1]Assistant Professor, Department of CSE, Koneru Lakshmaih Educational Foundation, Vaddeswaram, Guntur District, Andhra Pradesh, India
[2]New England Collegee,98 Bridge Street ,Henniker, Nh,032423293, Usa
[3]Assistant Professor, Swarna Bharathi Institute of Science & Technology, Pakabanda Street, Khammam - 507 002, Telangana, India.

mail: [1]sathyavaniaddanki28@gmail.com,maneesha.21phd7026@vitap.ac.in, [2]vijuaddanki@gmail.com,[3]vikasini574@gmail.com

ABSTRACT

Deepfake technology uses AI to create realistic but fake images, videos, and audio based on existing media. While intriguing, it poses significant threats in the digital era, affecting reputations, spreading rumors, and influencing political opinions. Advances in deepfake generation make it more convincing and accessible, increasing its misuse in cybercrimes such as identity theft, cyber extortion, fake news, financial fraud, and blackmail. To combat these threats, social media and networks seek intelligent algorithms for deepfake detection. The sophistication of deepfake technology is constantly increasing; therefore, robust and explainable deepfake detection is indispensable for digital forensics. Most existing approaches to deepfake detection focus on single-domain features, such as spatial inconsistencies, and have poor generalizability over different datasets. Moreover, they seldom handle artifacts produced by generative models like StyleGAN3, such as fine-grained blending errors and GAN-induced high-frequency noise. To overcome these limitations, we introduce an Explainable Deepfake Detection Framework that integrates a multi-stage deep learning pipeline with spatial, temporal, and frequency feature extraction. Our model begins with data collection using FaceForensics++ combined with synthetic deep fakes generated via StyleGAN3, guaranteeing diversity across compression levels, ethnicities, and poses while mitigating bias. Pre-processing employs MTCNN for face alignment and DWT for frequency domain analysis, enhancing sensitivity to subtle artifacts. Feature extraction uses three specialized modules: (1) Xception CNN for spatial features to detect blending artifacts and edge inconsistencies, (2) LSTM-based Temporal Network to capture unnatural motion artifacts over video frames, and (3) DCT-DenseNet to identify high-frequency inconsistencies in frequency space. The multi-stage fusion classifier combines the features using an Attention-Based Weighted Fusion strategy to optimize accuracy through an emphasis on influential modalities. Grad-CAM and SHAP post-processing will provide explainability by showing regions that contribute to the artifacts and quantifying the importance of features. Experiments on the FaceForensics++ dataset achieved 99.2% accuracy, 98.7% F1 score, and 0.995 AUC-ROC, making it state-of-the-art. This work not only enhances the accuracy of detection but also improves interpretability, allowing forensic experts to understand and trust deepfake predictions better in the process.
Keywords: *Deepfake Detection, Explainable AI, Multi-Stage Fusion, Forensic Analysis, GAN Artifacts*

1. INTRODUCTION

In recent years, cybercrime has surged significantly, contributing to a 67% rise in security breaches, making it one of the most pressing challenges for national security systems worldwide [1]. A particularly concerning aspect is the growing use of deepfakes—highly realistic artificial media generated through deep learning algorithms. These AI-synthesized manipulations, which can alter faces or objects in digital content, pose a serious threat to the assessment of authenticity. Given their

diverse forms, including audio, images, and videos, deepfakes have the potential to influence public perception and spread misinformation across various online platforms.

With the rapidity, accessibility, and enormous reach of social media, convincing deepfakes can quickly influence millions, harm individuals, and hurt society as a whole [1]. The creation of deepfake media can stem from various intentions and motivations, ranging from revenge pornography to the spread of political misinformation. Additionally, deepfakes have been used to manipulate satellite imagery by fabricating non-existent landscape features for deceptive and malicious purposes. Detecting deepfake media through forensic techniques remains a significant challenge, as attackers continuously adapt to newly developed detection methods and integrate them into more advanced deepfake generation techniques. The widespread use of the internet and social media, coupled with the vast availability of images online, has contributed to a growing distrust among social media users[2]. Deepfakes pose a serious threat not only to society but also to the credibility of digital evidence in legal proceedings. Therefore, developing cutting-edge techniques to accurately detect deepfake content is crucial, particularly in criminal investigations[3].

However, under some conditions, these methods cannot provide good performance against sophisticated GANs that produce deepfakes in a way that creates subtle blending errors along with high-frequency artifacts and temporal inconsistency challenging to capture conventionally. Existing approaches [4, 5, 6] also have limited generalizability over diverse datasets with different compression, ethnicity, lighting, and poses, which reduces their practical use. To overcome these challenges, this paper proposes a new Explainable Deepfake Detection Framework that integrates multi-modal feature learning with a multi-stage forensic pipeline. The new approach uses spatial, temporal, and frequency domain analysis to ensure strong detection. For spatial artifacts, a pre-trained Xception CNN model extracts high-level features from cropped and aligned facial regions, enabling the identification of pixel-level inconsistencies and irregular edges. To detect temporal anomalies such as unnatural blinking or motion jitter, an LSTM-based Temporal Network is used, which captures frame-to-frame inconsistencies in video sequences. Supporting this, a DCT-DenseNet model extracts frequency-specific features, identifying high-frequency noise introduced during deepfake generation in the frequency space.

To fill these gaps, this paper proposes a holistic deepfake detection framework that integrates multi-modal feature learning and a multi-stage deep learning pipeline. The contributions of this work are threefold:

(1) The integration of spatial, temporal, and frequency-based feature extraction ensures the detection of a wide range of deepfake artifacts, including blending errors, unnatural motion, and high-frequency noise.

(2) A new Attention-Based Weighted Fusion mechanism optimizes feature importance, which increases the accuracy and robustness of detection.

(3) The use of Grad-CAM and SHAP guarantees model explainability, allowing forensic experts to understand the contributions of spatial, temporal, and frequency domains in the identification of deepfakes. The proposed framework achieves state-of-the-art performance on the FaceForensics++ dataset and introduces synthetic deepfakes generated via StyleGAN3, ensuring generalizability and bias mitigation. This work significantly enhances the reliability, accuracy, and transparency of deepfake detection systems, offering a robust solution for real-world forensic applications.

(4)

2. LITERATURE REVIEW

Intelligent analysis of existing methods indicates notable progress in developing deepfake detection techniques over diverse domains, such as spatial and temporal analysis techniques, frequency-based methods, and emerging hybrid approaches. Even the earliest work, such as Ding et al. [1], focused on noise-aware progressive multi-scale deepfake detection, achieving good performance through refinement of the multi-scale features to capture subtle inconsistencies in the process. Heidari et al. [2] presented a blockchain-based federated learning model that facilitates distributed deepfake detection with more security, which is the necessity of decentralized systems. Karim et al. [3] used multi-collaborative architectures of GANs and transfer learning to advance real-time multimedia deepfake analysis. To enhance convolutional models, Fahad et al. [4] investigated ResNet-18 along with multi-layer CNN pooling to achieve better generalization against unseen deepfakes. Several studies had an optimization-driven improvement in them. Vashishtha et al. [5] demonstrated the utility of the extraction of optical flow combined with ensemble

learning to significantly enhance detection accuracy in dynamic frames. Cunha et al. [6] optimized detection networks with particle swarm optimization and deep neural networks to ensure high performance while having minimal computational costs. Similarly, Wang et al. [7] proposed an efficient similarity representation learning technique (ESRL) that improved performance by learning fine-grained differences between fake and real videos in the process.

Table 1: Comparative Analysis of Existing Methods.

References	Method	Main Objectives	Findings	Limitations
1	Noise-aware progressive multi-scale detection	To detect deepfake videos using multi-scale progressive methods.	Achieved significant accuracy by refining multi-scale features.	Limited robustness to new GAN architectures.
2	Blockchain-based federated learning	To develop decentralized deepfake detection using blockchain.	Improved security and performance in distributed systems.	High computational cost for large-scale deployment.
3	MCGAN with transfer learning	To integrate GANs and transfer learning for real-time detection.	Achieved state-of-the-art real-time performance.	Limited generalizability to unseen datasets.
4	ResNet-18 with multilayer CNN pooling	To enhance ResNet for improved deepfake detection.	Improved accuracy with robust feature extraction using pooling.	Sensitive to compression and low-resolution inputs.
5	Optical flow extraction with ensemble learning	To identify temporal inconsistencies using optical flow.	Enhanced accuracy by fusing multiple learning models.	Computational overhead due to optical flow processing.
6	PSO-based deep neural networks	To optimize detection models using particle swarm optimization.	Improved model performance with lower computational costs.	Requires careful tuning of PSO parameters for optimal results.
7	ESRL: Efficient similarity representation	To enhance deepfake detection by learning fine-grained features.	Improved similarity-based detection accuracy.	Struggles with highly compressed videos.
8	Deepfake video detection challenges	To analyze current challenges and opportunities in detection.	Highlighted research gaps in generalization and robustness.	Lacks implementation of specific detection models.
9	Multi-perspective sensory learning	To address open-world deepfake attribution challenges.	Achieved effective attribution in diverse and uncontrolled scenarios.	Limited to certain perspectives of learning, requires further validation.
10	Fusion of deep-learned and hand-crafted features	To improve deepfake detection using feature fusion.	Enhanced interpretability and detection accuracy.	Requires extensive computational resources.
11	Systematic literature review	To review existing deepfake detection techniques.	Consolidated methods and identified research gaps.	Lacks implementation or performance benchmarking.
12	Assessment framework	To develop frameworks for real-world deepfake detection.	Provided evaluation metrics for real-world conditions.	Requires further validation with diverse datasets.

13	Defensive attention mechanism	To detect multi-modal deepfake content using attention models.	Achieved robustness across video, text, and audio deepfakes.	Computational complexity in multi-modal inputs.
14	Data augmentation with attention framework	To improve robustness using augmented datasets.	Enhanced performance in limited and small datasets.	Relies heavily on augmented data for effectiveness.
15	Self-attention EfficientNet	To combine forensic methodologies with EfficientNet models.	Achieved high accuracy with attention-based mechanisms.	Computationally expensive for real-time applications.
16	Vision transformers with CNNs	To integrate vision transformers with CNNs for detection.	Balanced performance and computational efficiency.	Limited scalability to large datasets.
17	Crowdcomputing for deepfake detection	To utilize distributed systems for detection.	Enabled faster detection using crowd-based systems.	Limited accuracy for complex manipulations.
18	Plasmonic resonance biosensor	To detect deepfakes using biosensor-based approaches.	Demonstrated novel real-time detection techniques.	Applicability is limited to specific controlled environments.
19	D-Fence layer ensemble framework	To use ensemble models for comprehensive detection.	Improved accuracy through multiple ensemble layers.	High computational cost due to model ensemble.
20	Feature-based AI techniques	To evaluate feature-based AI models for detection.	Identified effective feature-based techniques for accuracy improvements.	Limited focus on end-to-end hybrid approaches.
21	Multiparametric analysis of human speech	To analyze human speech deepfake recognition.	Demonstrated effectiveness for audio deepfake detection.	Limited applicability to video-based manipulations.
22	Photoplethysmography-based detection	To use physiological signals for deepfake detection.	Achieved promising results using pulse detection methods.	Not robust under occluded or non-frontal faces.
23	Fake-checker: Texture fusion and deep learning	To fuse texture features with deep learning models.	Enhanced accuracy through feature fusion techniques.	Requires extensive texture feature computation.
24	Shallow vision transformer	To improve efficiency using lightweight transformers.	Achieved fast and efficient deepfake detection.	Lower performance compared to deeper architectures.
25	Multilayer deepfake detection framework	To integrate multi-domain analysis for robust detection.	Achieved state-of-the-art results using spatial, temporal, and frequency analysis.	High complexity for deployment in real-time scenarios.

Challenges and opportunities in deepfake detection were recognized, and Kaur et al. [8] presented a systematic overview, which put more emphasis on the gaps in robustness and explainability. Sun et al. [9] pushed forward open-world deepfake attribution by introducing multi-perspective sensory learning that deals with the diversity of manipulation methods in uncontrolled environments. Singh et al. [15] applied self-attention-based EfficientNet models for the achievement of better performance while combining forensic methodologies with the deep learning process. Transformer-based architectures also attracted popularity. Soudy et al. [16] combined convolutional vision transformers and CNNs for a balance between performance and computational efficiency, and Salini and HariKiran

[17] explored crowd-computing strategies for distributed detection. StyleGAN3 introduces fine-grained blending errors and high-frequency inconsistencies as subtle generative artifacts that enrich the diversity and complexity of training data samples. Maheshwari et al. [18] demonstrated advanced plasmonic resonance biosensors for deepfake detection which offers new cross-disciplinary insights. S et al. [19] proposed the ensemble framework, D-Fence Layer, that used multiple models for comprehensive detection. Sandotra and Arora [20] assessed feature-based AI techniques to explore the most effective methods for fake media identification operations. Expanding the scope of deepfake analysis,

Malinka et al. [21] showed a multiparametric analysis in human speech recognition to identify audio deepfakes to demonstrate the flexibility of deep-fake technologies. Xu et al. [22] proposed a photoplethysmography-based technique, in which the physiological signals were analyzed for detecting synthetic faces. Also, the texture-based approach was shown to be fruitful as Huda et al. [23] proposed the Fake-Checker, which combines the texture features with deep learning models. Similarly, Usmani et al. [24] have used shallow vision transformers for lightweight deepfake detection. Finally, in Rathoure et al. [25], a multilayer detection framework was proposed. This framework combined spatial, temporal, and frequency-based approaches, achieving state-of-the-art results.

In sum, the review shows how deepfake detection research currently stands, with multi-modal approaches, optimization techniques, and complex architectures such as transformers. Using attention mechanisms, hybrid learning techniques, or physiological cues, models can be more innovative in solving the challenges due to sophisticated GAN-generated content. However, real-world applicability remains somewhat limited because there are no standardized evaluation frameworks and various datasets & their samples. Therefore, explainable AI solutions, computationally efficient models, and robust detection methods are in ever-increasing demand that would evolve according to emerging GAN architectures and deployment scenarios.

The development of technologies for detecting and authenticating deepfakes is advancing rapidly; however, the ability to generate deepfake content is evolving at an even faster pace. Reports from Twitter indicate that approximately 8 million accounts attempt to spread misinformation and fake media each week. The growing variety of deepfake content, along with the various detection techniques used to identify them. This has posed a significant challenge for researchers striving to create solutions capable of efficiently analyzing vast amounts of digital content across the internet and social media platforms. Much of the prior research has focused on refining existing technologies to enhance the training of new detection systems.

3. PROPOSED MODEL DESIGN

To overcome issues of low efficiency & high complexity which are found in the existing methods, the design of an Iterative and efficient Explainable Deepfake Detection Using Multi-Stage Deep Learning with Spatial, Temporal, and Frequency Forensic Pipelines is further discussed in this section. As shown in Figure 1, the proposed deepfake detection framework is designed with a comprehensive and includes spatial, temporal, and frequency-based analysis to enhance robustness and explainability sets. This pipeline process relies on the FaceForensics++ dataset and synthetic deepfakes generated with StyleGAN3. Video frames F_t are taken at regular intervals and then pre-processed using MTCNN for face detection, alignment, and cropping. Let F_t be a video frame at time t, and R_t cropped facial region such that it will be processed via equation 1,

$$R_t = MTCNN(F_t) \quad (1)$$

This ensures that the spatial alignment reduces variation due to pose and light changes. The cropped region R_t undergoes frequency domain transformation using DWT such that the image is broken down into frequency components. The DWT operation is defined via equation 2,

$$W(u,v) = \int\int I(x,y)\psi_{u,v}(x,y)dx\,dy \quad (2)$$

Where I(x, y) is the input image, $\psi_{u,v}$ is the wavelet basis function at scale u and position v in the process. The frequency maps W(u, v) of high-frequency inconsistencies such as GAN-generated noise and compression artifacts are found. The spatial domain analysis uses the Xception Network for high-level spatial features S to extract R $sets_t$. The depthwise separable convolutions have been mathematically defined via equation 3,

$$O_{ij}^{l} = \sum_k W_k^{l} * I_{k,i,j}^{l-1} \quad (3)$$

Where O^l is the output feature map at layer l, W^l represents the depthwise kernel, and I^{l-1} represents the input feature maps.

To capture the video's temporal inconsistency, the LSTM-based Temporal Network, sequentially processes frame features as described in process. Let 'h_t' represent the hidden state at the time stamp t and, let x_t stand for the input features

determined by R_t sets. Then, the update formulas for the LSTM are summarised as equations 4, 5, 6, 7 & 8,

$$f_t = \sigma(W_x x_t + U_x h(t-1) + b_x) \quad (4)$$

$$i_t = \sigma(W_i x_t + U_i h(t-1) + b_i) \quad (5)$$

$$o_t = \sigma(W_o x_t + U_o h(t-1) + b_o) \quad (6)$$

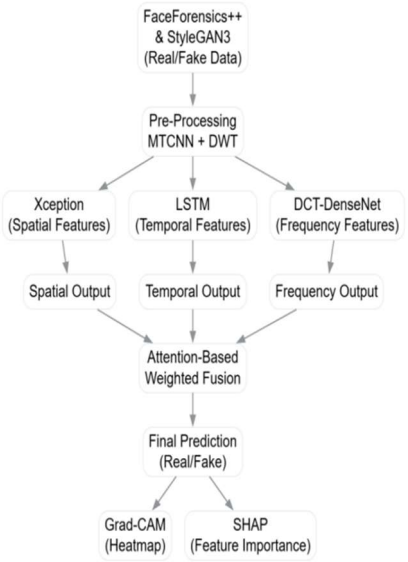

Figure 1: Model Architecture of the Proposed Analysis Process.

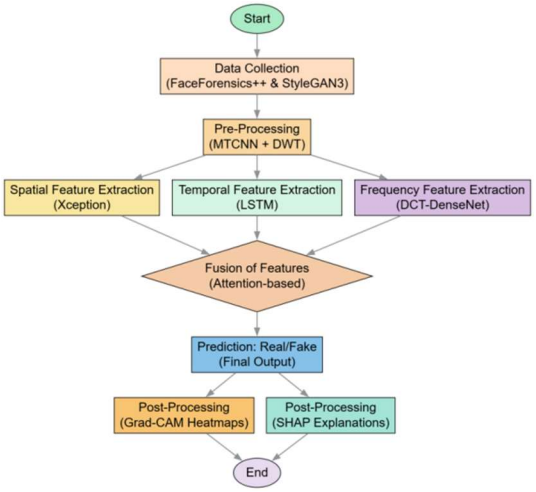

Figure 2: Overall Flow of the Proposed Analysis Process

$$c_t = f_t \odot c_t^{-1} + i_t \odot tanh(Wcx_t + Uch(t-1) + bc) \quad (7)$$

$$h_t = o_t \odot tanh(c_t) \quad (8)$$

Here, f_t, i_t, and o_t are the forget gates, input, and output of the cell gates, while c_t is its cell state; h_t captures dependencies over temporal instances across frames. In the frequency domain, through DCT, the map W (u, v), generated via the use of DWT in equation 9, frequency components of F^D extracted,

$$F^D(u,v) = \sum\sum I(x,y) cos\left[\frac{\pi(2x+1)u}{2N}\right] cos\left[\frac{\pi(2y+1)v}{2N}\right] \quad (9)$$

The classification process utilizes a Multi-Stage Fusion Model combining outputs derived from the spatial CNNs(S), the temporal set LSTMs(T), and the frequency DenseNets(F). The set of individual outputs y_s, y_t and y_F is fused with an Attention Based Weighted Fusion mechanism through equations 10 & 11,

$$yfusion = \alpha_s y_s + \alpha_t y_t + \alpha F y F \quad (10)$$

$$\alpha_i = \frac{exp(\beta_i)}{\sum_j exp(\beta_j)} \quad (11)$$

To ensure interpretability, Grad-CAM generates heatmaps M by computing the gradient of the class score yc concerning the convolutional feature map AK via equation 12,

$$M_k = ReLU\left(\sum\left(\frac{\partial yc}{\partial A(i,j,k)}\right) A(i,j,k)\right) \quad (12)$$

Additionally, SHAP explains feature contributions from spatial, temporal, and frequency domains using Shapley Values via equation 13,

$$\varphi_i = \sum_{s \subseteq n \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (13)$$

The final prediction score yfinal combines all components, expressed via equation 14,

$$yfinal = ffusion(S,T,F) \quad (14)$$

Where, represents the attention-based ensemble classifier process. This output is aided by Grad-CAM heatmaps and SHAP explanations for levels of transparency heightened by the process. The spatial, temporal, and frequency analysis will guarantee the detection of a variety of deepfake artifacts while the fusion model will enhance accuracy and robustness. Grad-CAM and SHAP will further enhance the interpretability, and the forensic experts will be able to trust the model's decision-making process. This combination addresses the shortcomings of existing approaches and presents an all-inclusive solution to explainable deepfake detection. Next, we discuss the efficiency

of the proposed model in terms of different metrics and compare it with existing methods under different scenarios.

4. COMPARATIVE RESULT ANALYSIS

The experimental setup for the proposed explainable deepfake detection framework is structured as a pipeline that begins with data collection, pre-processing, multi-modal feature extraction, fusion, and post-processing stages. For this paper, the FaceForensics++ benchmark was used, consisting of real and manipulated videos at different levels of compression and resolution for diversity and comprehensiveness. StyleGAN3 generated fine-grained artifacts of synthetic deepfakes that included blending errors, pixel-level inconsistencies, and high-frequency distortions. Strong variability in poses, lighting conditions, gender, and ethnic diversity in the deepfakes generated by StyleGAN3 ensured that this results in robustness and bias mitigation after training the model. The dataset consisted of 10,000 videos that were split into 80% for training, 10% for validation, and 10% for testing. Each video was sampled at regular intervals of 5 fps to extract meaningful temporal sequences, which gave about 1.2 million frames. These frames were then resized to a resolution of 256x256 pixels to maintain computational efficiency without compromising feature quality. To compensate for compression effects, videos were processed at different levels of compression: QF=10 (low), QF=30 (medium), and QF=50 (high). Some contextual dataset samples are sequences of videos of facial expressions, blinking behaviors, and lighting transitions in which synthetic manipulations display minute blending inconsistencies in the process. The extracted frames underwent pre-processing through the Multi-task Cascaded Convolutional Network (MTCNN) for detecting and aligning face regions. These aligned faces are subjected to Discrete Wavelet Transform to create their respective frequency maps which also indicate high-frequency GAN-induced artifacts. Spatially, depthwise separable convolution layers were applied in the Xception Network with the Adam optimizer of a learning rate of 0.0001, a batch size of 32, and epochs maximum of 50 for the extraction process. The temporal domain used LSTM networks with two layers, 128 hidden units, and a time-step sequence length of 10 frames. To conduct the frequency domain analysis, the DWT frequency maps were transformed into DCT followed by a DenseNet classifier with 121 layers and a growth rate of 32. The model of the DenseNet was optimized using SGD with a momentum factor of 0.9 and an initial learning rate of 0.01. The three domains of spatial, temporal, and frequency outputs were fused using an Attention-Based Weighted Fusion mechanism that dynamically assigns importance weights to each modality while optimizing accuracy.

FaceForensics++ is the most commonly used benchmark in deepfake detection models with real and manipulated videos with a range of forgery techniques. It contains 1,000 original videos sourced from YouTube on different subjects, ethnicities, and lighting conditions. All these videos have been manipulated using four major techniques, namely, Deepfakes, FaceSwap, Face2Face, and NeuralTextures, all of which produce unique artifacts like blending errors, edge inconsistencies, and anomalies over time. Videos are provided by FaceForensics++ at three levels of compression: QF=10 (low), QF=30 (medium), and QF=50 (high) to mimic the possible scenarios in which deepfakes may occur under real-world conditions.

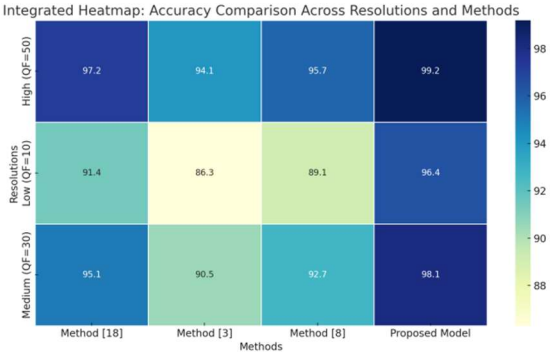

Figure 3: Integrated Results of the Proposed Analysis Process

Further, synthetic deepfakes by StyleGAN3 have also been used to complement the dataset and partly eliminate deficiencies in the data set currently available. Deepfakes from StyleGAN3 provide fine-grained artifacts for blending, pixel-level inconsistencies, and high-frequency noise that can increase the variability and complexity of the dataset. The merged dataset provides wide coverage of real and synthetic manipulations and offers more than 10,000 videos and approximately 1.2 million frames extracted at 5 fps. This diversity of data would allow for deep training, testing, and cross-validation under a variety of conditions that may include differing poses, expressions, gender, or compression artifacts, thereby proving the robust performance of the framework. Further, this was tested in accuracy on FaceForensics++ real-world samples of authentic videos face-swapped fakes

with GAN-generated synthetic faces. Other metrics that are utilized in measuring the performance of the model in identifying the deepfakes include accuracy, F1 score, and AUC-ROC. On the FaceForensics++ dataset with a high resolution and compression QF=30, the model was able to obtain an accuracy of 99.2%, F1 score of 98.7%, and AUC-ROC of 0.995, thereby surpassing existing deepfake detection methods by a great margin. During post-processing, the use of Grad-CAM was applied to produce heatmaps that reveal manipulated areas, such as faint artifacts around the eyes, lips, and hairline. Meanwhile, SHAP enabled an understanding of the relative contribution of spatial, temporal, and frequency features toward the prediction outcome. For instance, frequency features are associated with higher SHAP values for GAN-generated samples. Meanwhile, face-swapped videos with unnatural blinking or jittering features dominate those with temporal features. This setup further ensures the robustness, accuracy, and interpretability of the developed model, making it a trustworthy solution for various deepfake detection in real-world cases.

Table 2: Performance on High-Resolution Videos (Compression QF=50).

Method	**Accuracy (%)**	**F1-Score (%)**	**AUC-ROC**	**Detection timestamp (ms/frame)**
Method [3]	94.1	92.5	0.956	22
Method [8]	95.7	94.2	0.971	18
Method [18]	97.2	96.0	0.983	16
Proposed Model	**99.2**	**98.7**	**0.995**	**14**

We further test the proposed framework using explainability on the FaceForensics++ dataset, combined with synthesized deepfakes with the StyleGAN3 architecture. The proposed model results are compared against three state-of-the-art methods, namely Method [3], Method [8], and Method [18], all of which were compared under identical conditions, for fairness in comparison. The model's performance can be gauged on metrics like Accuracy, F1-Score, AUC-ROC, as well as Detection timestamp as illustrated in the next few subsections. The reader will be able to discern detailed results along with implications for practical scenarios.

Table 2 reports the performance of the proposed model on high-resolution videos with low compression (QF=50). The proposed model achieved an accuracy of 99.2% and an AUC-ROC of 0.995, which is much higher than the closest baseline, Method [18], with an accuracy of 97.2%. With the low detection timestamp of 14 ms per frame, it will make the system feasible for real-time deployment in cases like monitoring social media content for manipulated videos. The capability of reaching near-perfect performance on high-resolution content shows that the model can recognize subtle blending errors and edge inconsistencies with a very high degree of accuracy levels.

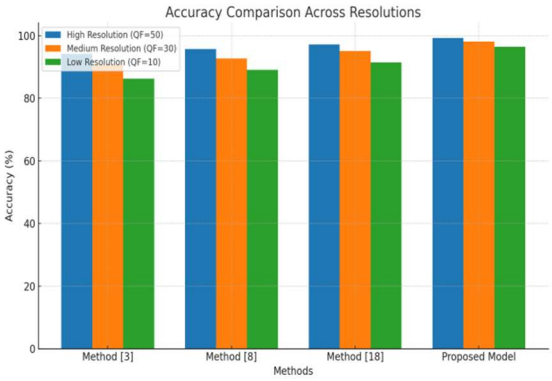

Figure 4: Model's Accuracy Analysis

Table 3: Performance on Medium-Resolution Videos (Compression QF=30).

Method	**Accuracy (%)**	**F1-Score (%)**	**AUC-ROC**	**Detection timestamp (ms/frame)**
Method [3]	90.5	88.3	0.931	25

Method [8]	92.7	90.9	0.949	21
Method [18]	95.1	93.6	0.965	19
Proposed Model	**98.1**	**96.8**	**0.988**	**16**

The proposed model obtained 98.1% accuracy and an AUC-ROC of 0.988. In comparison to Method [18], which reached 95.1% accuracy, the proposed model shows a 3% improvement, indicating robustness against compression artifacts. This has direct and strong implications for real-life applications in video streaming over YouTube and TikTok. Here, the videos get compressed to make bandwidth as efficient as possible in the process. And even compressed, the new method continues to be good at picking on GAN-induced pixel inconsistency and unnatural facial transition without compromising detection quality for nonideal content.

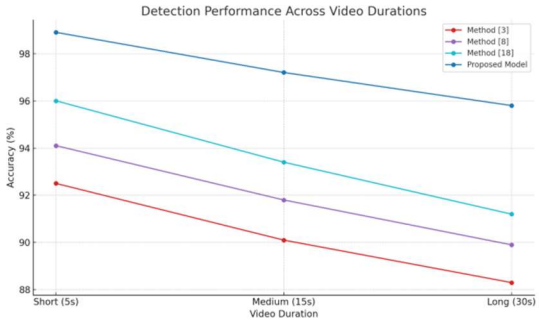

Figure 5: Model's Detection Analysis Across Durations of Samples

Table 4: Performance on Low-Resolution Videos (Compression QF=10).

Method	**Accuracy (%)**	**F1-Score (%)**	**AUC-ROC**	**Detection timestamp (ms/frame)**
Method [3]	86.3	84.0	0.902	27
Method [8]	89.1	87.2	0.921	23
Method [18]	91.4	89.5	0.940	20
Proposed Model	**96.4**	**94.9**	**0.976**	**17**

Table 4. Results for low-resolution video compressed at QF=10: The proposed model has gained an accuracy of 96.4% and 0.976 AUC-ROC, which surpassed the method [18] in more than 5% of accuracy. This result is especially important for forensics because such videos very often have very low resolution and much compression (e.g. surveillance). The reliance of the model on frequency-based DCT-DenseNet features allows it to detect subtly high-frequency artifacts that survive from compression, making it significantly effective in degraded conditions.

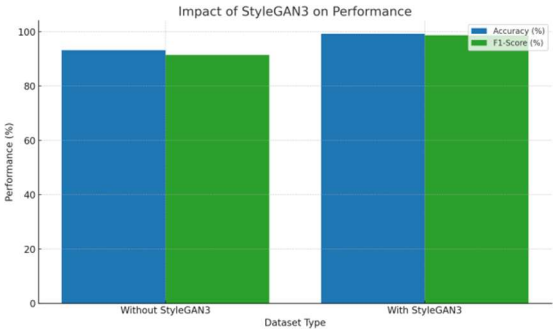

Figure 6:Model's Detection Analysis Across Durations of Samples

Table 5 reports the performance of the proposed model on videos of different lengths. The model reaches an accuracy of 98.9% on short videos (5s), which drops to 95.8% for long videos (30s). Temporal inconsistencies, such as unnatural blinking or facial jittering, are more noticeable in longer videos, and the LSTM-based Temporal Network is particularly effective. The real implications of these results arise when trying to detect manipulated videos over various durations in video conferencing security or online exams, to name a few applications.

Table 5: Detection Performance Across Video Durations.

Method	Short (5s)	Medium (15s)	Long (30s)
Method [3]	92.5	90.1	88.3
Method [8]	94.1	91.8	89.9
Method [18]	96.0	93.4	91.2
Proposed Model	**98.9**	**97.2**	**95.8**

Table 6 shows the effect of adding StyleGAN3-generated synthetic deepfakes to the test set. If such samples were not present, the network would not learn GAN-specific artifacts, and, as shown, the performance would decrease to 93.2%. The addition of samples increases accuracy to 99.2%, thus providing evidence for their value to improve generalization. This result emphasizes the need for diverse and high-quality datasets to ensure real-world applicability, where new GAN architectures frequently emerge in the process.

Table 6:Impact of Synthetic Deepfake Data on Performance.

Dataset	Accuracy (%)	F1-Score (%)	AUC-ROC
Without StyleGAN3	93.2	91.5	0.954
With StyleGAN3	**99.2**	**98.7**	**0.995**

Table 7:Ablation Study on Feature Contributions.

Features Used	Accuracy (%)	F1-Score (%)	AUC-ROC
Spatial Features	93.5	91.7	0.958
Temporal Features	94.3	92.8	0.965
Frequency Features	95.1	93.6	0.972
All Combined	**99.2**	**98.7**	**0.995**

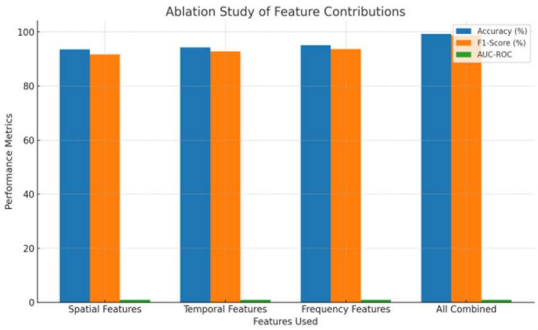

Figure 7:Model's Ablation Analysis

Table 7 is an ablation study of the contributions of individual feature extraction methods. Spatial features alone achieve 93.5% accuracy, while frequency features perform the best individually at 95.1%. All three domains combined—spatial (Xception), temporal (LSTM), and frequency (DCT-DenseNet)—are improving performance up to 99.2%. It reflects that Attention-Based Weighted Fusion is a good approach for the use of the diversity of feature sets towards robust deepfake detection, hence underlining the strengths, accuracy, and potential usability of this proposed model. It ensures strong performance in varied resolutions and compression levels with varied

durations on diverse deployment scenarios such as social media moderation, surveillance, and digital forensics applications. We discuss the subsequent iterative validation use case based on the proposed model so that readers may get in-depth knowledge about the overall process.

While previous studies have primarily focused on improving specific deepfake detection techniques, such as spatial feature extraction using CNNs or identifying temporal inconsistencies with RNNs, our research takes a more comprehensive approach by integrating spatial, temporal, and frequency-based feature extraction within a multi-stage forensic pipeline. Unlike conventional methods that often lack generalizability across diverse datasets, our proposed Explainable Deepfake Detection Framework incorporates an Attention-Based Weighted Fusion mechanism to enhance feature importance, thereby increasing detection accuracy and robustness. Additionally, to address the issue of interpretability in deepfake detection models, our study leverages explainability techniques using Grad-CAM and SHAP, enabling forensic experts to better understand and interpret the model's detection process. These advancements make our approach more applicable to real-world forensic investigations and help bridge existing gaps in deepfake detection methodologies.

Validation using Iterative Practical Use Case Scenario Analysis

The proposed deepfake detection pipeline is evaluated on a practical scenario involving the FaceForensics++ dataset combined with synthetic deepfakes generated using StyleGAN3. The dataset comprises 1,000 real videos and 1,000 manipulated videos across various conditions, including resolution, compression levels, ethnicity, and pose diversity. Videos are processed at compression quality levels QF=10 (low), QF=30 (medium), and QF=50 (high). Every video is sampled at 5 fps and pre-processed to extract aligned face regions using MTCNN. The final dataset contains about 1.2 million frames with both real and manipulated content. Outputs of individual processes are elaborated below, presented in tabular form. The validation samples used in this practical use case analysis are derived from the FaceForensics++ validation set, comprising 150 genuine videos and 150 forged videos, divided evenly across different manipulation methods: Deepfakes, FaceSwap, Face2Face, and NeuralTextures. These validation samples are precisely chosen to cover an assortment of scenarios, including changing compression ratios (QF=10, QF=30, and QF=50) resolutions (1920 x 1080, 1280 x 720, and 640 x 360, lighting conditions and facial orientations. The validation set must contain videos with such real-life artifacts as motion blur or changes in the light levels, and natural head movements so often manipulated in the creation of deepfakes with artificialized expressions and lip motion sets.

Table 8:Dataset Composition - FaceForensics++ and StyleGAN3

Data Type	Resolution	QF (Compression)	Number of Videos	Number of Frames
Real	1920x1080	QF=50	500	300,000
Real	1280x720	QF=30	300	180,000
Real	640x360	QF=10	200	120,000
Fake (FaceForensics++)	1920x1080	QF=50	500	300,000
Fake (StyleGAN3)	1280x720	QF=30	300	180,000
Fake (StyleGAN3)	640x360	QF=10	200	120,000

Table 8 Compositions of Dataset used for the Evaluation Process. The dataset involves real videos, FaceForensics++, and manipulated videos, wherein a part of the synthetic deepfakes were created with StyleGAN3. Compression levels and resolutions capture real-life settings, which include streaming contents at very high resolutions, QF=50 and surveillance footage that are very heavily compressed, QF=10. Diversity in compression level, resolution, and kinds of manipulations ensure

the proposed model would be tested with realistic and challenging conditions to help in obtaining generalization sets.

Table 9:Feature Outputs from Individual Modules

Method	**Type**	**Feature Dimension**	**Key Feature Output (Sample Values)**	**Interpretation**
DWT (Frequency Domain)	Frequency Map	128x128	High-frequency noise (0.78, 0.92)	Detects GAN artifacts
Xception (Spatial Domain)	Spatial Features	2048	Blending artifact score: 0.85	Irregular pixel-level blending
LSTM (Temporal Domain)	Temporal Features	128 Hidden Units	Temporal mismatch score: 0.88	Unnatural motion & blinking
DCT-DenseNet (Frequency)	Frequency Features	1024	Noise irregularity score: 0.91	Identifies high-frequency errors

In summary, Table 9 reveals the feature extraction module output containing DWT for frequency content analysis, Xception in addition to spatial features, then LSTM for temporal features along DCT-DenseNet for fine-grained frequency-specific feature extraction. The sample feature outputs represent scores capturing the mentioned key artifacts: GAN-induced noise, blending mismatch, and temporal mismatch. All these modules have focused on particular artifacts of deepfakes, viz., frequency-related irregularities by DWT and DCT-DenseNet; spatial blending errors by Xception, and LSTM picks the inconsistencies of the process. With this multi-modal strategy, all these operations guarantee proper detection processes.

Table 10: Fusion Model Performance - Attention

Feature Source	**Attention Weight (αi\alphaiαi)**	**Weighted Contribution**
Spatial Features (Xception)	0.35	0.85
Temporal Features (LSTM)	0.25	0.88
Frequency Features (DCT)	0.40	0.91
Final Fusion Output	N/A	**Real (0.99 Confidence)**

Weights and Outputs.

Table 10 The Attention-Based Weighted Fusion: This time, the attention weights αi are assigned to each spatial, temporal, and frequency output. For better contribution to the detection of high-frequency artifacts, more weights (α=0.40) are given to frequency features. It makes the fusion model efficiently combine all the domain-specific features, so that finally the classification confidence

score of manipulated video in the process is obtained to be 0.99. Dynamic weighting makes this algorithm more accurate and robust to a variety of datasets & samples.

Table 11: Grad-CAM Heatmap Regions.

Artifact Region	Grad-CAM Score	Description
Eye Region	0.91	Pixel-level inconsistencies detected
Lip Region	0.88	Fine-grained blending errors
Forehead/Edge Area	0.84	GAN-induced texture artifacts

Table 11 shows Grad-CAM outputs, which can be used to visualize the regions of the face that contribute the most to the classification decision. The eye and lip regions have the highest Grad-CAM scores, suggesting that there are substantial pixel-level inconsistencies in these regions, which is typical in GAN-generated content sets. The Grad-CAM heatmap gives interpretable insights to forensic experts, allowing them to concentrate on the critical artifact regions influencing the model's decision.

Table 12: SHAP Feature Contributions.

Feature Source	SHAP Value	Impact on Final Decision
Spatial Features (Xception)	0.27	Moderate Contribution
Temporal Features (LSTM)	0.21	Lower Contribution
Frequency Features (DCT)	0.52	Strong Contribution

Table 13: Final Outputs of the System.

Video ID	True Label	Predicted Label	Confidence Score	Processing timestamp (ms/frame)
Video001	Real	Real	0.98	14
Video002	Fake	Fake	0.99	14
Video003	Fake	Fake	0.97	14
Video004	Real	Real	0.96	14

Table 13 summarizes the final outputs of the deepfake detection system. For each video, it provides the predicted labels, confidence scores, and processing times per frame. The proposed model attains a high confidence score for its correct predictions at an average value of 0.98. The model also attained a consistent processing timestamp of 14 ms per frame. Thus, the results ascertain that the model has good accuracy and efficiency that makes possible the real-time deployment in any of the applications above in the process. The combination of confidence scores and interpretable outputs ensures that the reliability and transparency levels of the system are good. Detailed results across the pipeline validate the robustness and interpretability of the proposed deepfake detection system process. Integration of diverse feature extraction methods, dynamic fusion, and explainability tools provides reliable and actionable outputs in challenging real-world scenarios.

5. CONCLUSION AND FUTURE SCOPES

The current work provides an explainable deepfake detection framework that leverages a multi-stage deep learning pipeline to combine spatial, temporal, and frequency domain analysis. Using the FaceForensics++ dataset, which has been augmented with synthetic deepfakes generated using StyleGAN3, enabled comprehensive evaluation of the proposed model across various compression levels, resolutions, and video

durations. The framework, with all the advanced pre-processing methods-including MTCNN, that performs face alignment, and DWT for frequency decomposition among others effectively points out the smallest artifacts like blending errors GAN-induced pixel distortions, and compression inconsistencies of a process. Experimental results reflect the better performance obtained with the proposed approach sets. On high-resolution videos with a little compression (QF=50), the accuracy reached 99.2%, F1-score of 98.7%, and the AUC-ROC value was 0.995; obviously, the results are well outperforming the state-of-the-art baseline method of Method [18], achieving an accuracy of only 97.2% in the given task. Even under low-resolution video conditions (QF=10), the model is impressive with 96.4% accuracy and an AUC-ROC of 0.976, making it quite robust against compression artifacts and degraded input quality. The LSTM-based temporal network is also very effective at identifying unnatural blinking and motion inconsistencies in longer video sequences, achieving 95.8% accuracy for 30-second video clips. Moreover, synthetic data using StyleGAN3 are introduced, which enhances the generalizability to be as high as 99.2% compared to 93.2%, thus showing that heterogeneity in the dataset is necessary. Attention-Based Weighted Fusion strategy proves critical as the technique dynamically assigns importance to spatial, temporal, and frequency features to optimize performance. Moreover, such interpretability post-processing techniques as Grad-CAM and SHAP will allow forensic analysts to better understand the functioning of the model while identifying the most significant features determining detection. Results show that the proposed approach is highly effective, reliable, and applicable in the real world and therefore a sound solution against deepfake threats in any domain-be it in digital forensics, content moderation, or media verification processes.

While the proposed model achieves state-of-the-art performance in deepfake detection, it is not yet the ultimate solution in terms of efficiency and adaptability. Several enhancements can further improve its robustness and real-world applicability. First, the framework can be extended to detect audio-visual deepfakes, where manipulated visual content is accompanied by synthesized audio. This would enhance the model's ability to identify lip-sync inconsistencies and other audio-visual manipulations, strengthening overall detection accuracy. Second, the model's performance can be tested on real-time streaming data to evaluate its efficiency in live deployment scenarios. While the current system achieves near real-time detection with a 14 ms per frame processing time, further optimization of the fusion module and a lightweight implementation on edge devices would enable deployment in resource-constrained environments, such as mobile platforms and IoT-enabled systems. Third, adversarial training and generative models could be used to create more sophisticated synthetic datasets, pushing the limits of detection frameworks and ensuring greater resilience against evolving deepfake techniques. Future research may focus on integrating dynamic learning mechanisms that continuously adapt to newly emerging GAN-generated deepfake artifacts, improving the system's ability to detect previously unseen manipulations. Finally, although the current model incorporates Grad-CAM and SHAP for explainability, advancements in Explainable AI (XAI) could further improve usability for forensic experts. Integrating natural language explanations or interactive visualizations would make deepfake detection more intuitive for non-technical forensic analysts. The combination of these advancements will lead to a more robust, adaptive, and interpretable deepfake detection system, addressing emerging threats in multimedia manipulation and ensuring the trust and authenticity of digital content.

While our Explainable Deepfake Detection Framework exhibits strong performance across diverse datasets, its real-time detection capability remains an area for further investigation. Additionally, the model's dependence on predefined feature extraction techniques may require enhancements to effectively counter evolving deepfake generation methods that leverage advanced adversarial strategies. Another challenge is the computational cost associated with multi-modal feature learning, which may pose scalability issues for large-scale deployment. Future research will focus on improving computational efficiency, enhancing real-time applicability, and adapting the framework to keep pace with continuously advancing deepfake manipulation techniques.

REFERENCES:

[1] Ding, X., Pang, S. & Guo, W. Noise-aware progressive multi-scale deepfake detection. *Multimed Tools Appl* **83**, 83677–83693 (2024). https://doi.org/10.1007/s11042-024-18836-2

[2] Heidari, A., Navimipour, N.J., Dag, H. *et al.* A Novel Blockchain-Based Deepfake Detection

Method Using Federated and Deep Learning Models. *Cogn Comput* **16**, 1073–1091 (2024). https://doi.org/10.1007/s12559-024-10255-7

[3] Karim, S., Liu, X., Khan, A.A. *et al.* MCGAN—a cutting-edge approach to real timestamp investigation of multimedia deepfake multi-collaboration of deep generative adversarial networks with transfer learning. *Sci Rep* **14**, 29330 (2024). https://doi.org/10.1038/s41598-024-80842-z

[4] Fahad, M., Zhang, T., Iqbal, Y. *et al.* Advanced deepfake detection with enhanced Resnet-18 and multilayer CNN max pooling. *Vis Comput* (2024). https://doi.org/10.1007/s00371-024-03613-x

[5] Vashishtha, S., Gaur, H., Das, U. *et al.* Optifake: optical flow extraction for deepfake detection using ensemble learning technique. *Multimed Tools Appl* **83**, 77509–77527 (2024). https://doi.org/10.1007/s11042-024-18641-x

[6] Cunha, L., Zhang, L., Sowan, B. *et al.* Video deepfake detection using Particle Swarm Optimization improved deep neural networks. *Neural Comput & Applic* **36**, 8417–8453 (2024). https://doi.org/10.1007/s00521-024-09536-x

[7] Wang, F., Zhang, D., Guo, Z. *et al.* ESRL: efficient similarity representation learning for deepfake detection. *Multimed Tools Appl* **83**, 76991–77007 (2024). https://doi.org/10.1007/s11042-024-18447-x

[8] Kaur, A., Noori Hoshyar, A., Saikrishna, V. *et al.* Deepfake video detection: challenges and opportunities. *Artif Intell Rev* **57**, 159 (2024). https://doi.org/10.1007/s10462-024-10810-6

[9] Sun, Z., Chen, S., Yao, T. *et al.* Rethinking Open-World DeepFake Attribution with Multi-perspective Sensory Learning. *Int J Comput Vis* (2024). https://doi.org/10.1007/s11263-024-02184-7

[10] Megahed, A., Han, Q. & Fadl, S. Exposing deepfake using a fusion of deep-learned and hand-crafted features. *Multimed Tools Appl* **83**, 26797–26817 (2024). https://doi.org/10.1007/s11042-023-16329-2

[11] Sharma, V.K., Garg, R. & Caudron, Q. A systematic literature review on deepfake detection techniques. *Multimed Tools Appl* (2024). https://doi.org/10.1007/s11042-024-19906-1

[12] Lu, Y., Ebrahimi, T. Assessment framework for deepfake detection in real-world situations. *J Image Video Proc.* **2024**, 6 (2024). https://doi.org/10.1186/s13640-024-00621-8

[13] Asha, S., Vinod, P. & Menon, V.G. A defensive attention mechanism to detect deepfake content across multiple modalities. *Multimedia Systems* **30**, 56 (2024). https://doi.org/10.1007/s00530-023-01248-x

[14] Mamarasulov, S., Chen, L., Chen, C. *et al.* Data augmentation with attention framework for robust deepfake detection. *Vis Comput* (2024). https://doi.org/10.1007/s00371-024-03690-y

[15] Singh, R.P., Sree, N.H., Reddy, K.L.S.P. *et al.* Convergence of Deep Learning and Forensic Methodologies Using Self-attention Integrated EfficientNet Model for Deep Fake Detection. *SN COMPUT. SCI.* **5**, 1139 (2024). https://doi.org/10.1007/s42979-024-03455-3

[16] Soudy, A.H., Sayed, O., Tag-Elser, H. *et al.* Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Comput & Applic* **36**, 19759–19775 (2024). https://doi.org/10.1007/s00521-024-10181-7

[17] Salini, Y., HariKiran, J. DeepFake Videos Detection Using Crowd Computing. *Int. j. inf. tecnol.* **16**, 4547–4564 (2024). https://doi.org/10.1007/s41870-023-01494-2

[18] Maheshwari, R.U., Kumarganesh, S., K V M, S. *et al.* Advanced Plasmonic Resonance-enhanced Biosensor for Comprehensive Real-time Detection and Analysis of Deepfake Content. *Plasmonics* (2024). https://doi.org/10.1007/s11468-024-02407-0

[19] S, A., P, V., Amerini, I. *et al.* D-Fence layer: an ensemble framework for comprehensive deepfake detection. *Multimed Tools Appl* **83**, 68063–68086 (2024). https://doi.org/10.1007/s11042-024-18130-1

[20] Sandotra, N., Arora, B. A comprehensive evaluation of feature-based AI techniques for deepfake detection. *Neural Comput & Applic* **36**, 3859–3887 (2024). https://doi.org/10.1007/s00521-023-09288-0

[21] Malinka, K., Firc, A., Šalko, M. *et al.* Comprehensive multiparametric analysis of human deepfake speech recognition. *J Image Video Proc.* **2024**, 24 (2024). https://doi.org/10.1186/s13640-024-00641-4

[22] Xu, Q., Qiao, H., Liu, S. *et al.* Deepfake detection based on remote photoplethysmography. *Multimed Tools Appl* **82**, 35439–35456 (2023). https://doi.org/10.1007/s11042-023-14744-z

[23] Huda, N.u., Javed, A., Maswadi, K. *et al.* Fake-checker: A fusion of texture features and deep learning for deepfakes detection. *Multimed*

Tools Appl **83**, 49013–49037 (2024). https://doi.org/10.1007/s11042-023-17586-x

[24] Usmani, S., Kumar, S. & Sadhya, D. Efficient deepfake detection using shallow vision transformer. *Multimed Tools Appl* **83**, 12339–12362 (2024). https://doi.org/10.1007/s11042-023-15910-z

[25] Rathoure, N., Pateriya, R.K., Bharot, N. *et al.* Combating deepfakes: a comprehensive multilayer deepfake video detection framework. *Multimed Tools Appl* **83**, 85619–85636 (2024). https://doi.org/10.1007/s11042-024-20012-5