International Journal for Electronic Crime Investigation

ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2025.0902/264

Research Article

Vol. 9 issue 2 Jul-Dec 2025

Forensic Lens: Deepfake Detection Through Micro-Level Facial Blood-Flow Signals

Hassan Minhal Raza [1], Mahnoor Ahmad [1], Nadeem Jabbar [2], Sanya Abdullah [2]

[1]Department of Cyber Security, The Superior University Lahore, Pakistan

[2]Faculty of Computer Science & Information Technology, The Superior University Lahore, Pakistan

Corresponding Author: hassanminhalraza@gmail.com

Received: Dec 9,2025, **Accepted:** Dec 18,2025; **Published:** Dec 30,2025

ABSTRACT

Deepfake technology, propelled by recent advances in deep learning and most notably by generative adversarial networks (GANs), has evolved far beyond its early applications in entertainment. What began as a tool for playful visual aug- mentation has now emerged as a substantive challenge to digital privacy, information integrity, and public trust. As synthetic media approaches near-photorealistic fidelity, traditional detec- tion strategies—whether based on conspicuous visual artifacts or computationally intensive convolutional neural networks—are increasingly strained. In practice, these methods often reveal shortcomings in scalability, exhibit sensitivity to dataset bias, and demand prohibitive computational resources, making them difficult to deploy in real-world scenarios. To address these limita- tions, this study introduces Forensic Lens, a lightweight deepfake detection framework that shifts focus from appearance-centric analysis to physiological consistency. The approach leverages remote photoplethysmography (rPPG) signals, capturing im- perceptible facial color fluctuations induced by cardiovascular activity. These signals are then embedded within a similarity graph, enabling semi-supervised label propagation across both annotated and unannotated samples. By grounding detection in intrinsic physiological cues rather than purely visual patterns, the framework improves generalization while reducing reliance on large, exhaustively labeled datasets. Extensive experiments con- ducted on the Celeb-DF v2 benchmark demonstrate that Forensic Lens achieves an accuracy of 90%, comparable to contemporary CNN-based detectors yet attained with markedly lower compu- tational overhead. Beyond quantitative performance, the model offers interpretability and resilience against compression artifacts and noise—qualities often overlooked but essential in forensic practice. These characteristics make the framework particularly well suited for deployment on resource-constrained platforms, including mobile devices and browser-based monitoring tools, where efficiency and reliability are paramount.

Keywords: Deepfake, Remote photoplethysmography rPPG, Semi-supervised, Lightweight detection framework, Cyberse- curity, CNN

1. INTRODUCTION

Deepfake technology, propelled by advances in deep learning and particularly by generative adversarial networks (GANs), has enabled the synthesis of strikingly realistic images and videos. By learning latent representations from authentic media and transferring them to fabricated con- tent, these systems produce synthetic outputs that are often indistinguishable from reality. Initially, such techniques were embraced within entertainment and creative ecosystems—including platforms such as Snapchat, TikTok, and Bigo—for benign purposes such as visual augmentation and storytelling [50]. Yet, the trajectory of deepfakes has quickly shifted from playful experimentation to a profound threat to digital privacy, security, and societal trust. Malicious exploita- tion now spans political manipulation, media impersonation, cybersecurity breaches, identity theft, and large-scale financial fraud [6], [52]. A striking example occurred in 2024, when a convincingly generated deepfake video of Elon Musk was disseminated to promote a cryptocurrency scam, resulting in substantial financial losses [51]. Such incidents underscore both the sophistication of synthetic media and the urgent need for reliable forensic countermeasures.

Within the research community, deepfakes are increas- ingly regarded as a critical challenge due to their abil- ity to erode public confidence in digital media, amplify misinformation, and compromise individual safety [9], [50]. Early detection strategies largely relied on identifying visual artifacts or training convolutional neural network (CNN) classifiers on appearance-based cues. However, contemporary GAN-generated content has effectively neutralized many of these telltale signs, suppressing irregularities in blinking, facial symmetry, or lighting [53]. As a result, CNN-centric solu- tions often exhibit limited robustness when confronted with cross-dataset variations, aggressive compression, or real-world noise. Their dependence on large annotated datasets and computationally demanding architectures further constrains scalability and deployment in resource-limited environments [1], [17]. These technical limitations are compounded by eth- ical concerns surrounding privacy, consent, and the potential misuse of forensic technologies themselves [30].

Recent research has shifted toward physiological signal analysis, motivated by the observation that authentic videos inherently preserve subtle biological rhythms that generative models struggle to reproduce faithfully. Remote photoplethys-mography (rPPG), in particular, captures minute facial color variations induced by cardiovascular activity, offering a biolog- ically grounded signal for authenticity verification. Pioneering works such as DeepRhythm and FakeCatcher demonstrated that rPPG-based representations provide discriminative fea- tures that remain largely invariant to visual realism [10], [43]. Despite their promise, existing rPPG-driven approaches remain sensitive to compression, sensor noise, and dataset diversity, limiting their effectiveness in unconstrained settings. Surveys in media forensics [39] have highlighted these shortcomings and emphasized the need for lightweight, interpretable detec- tion frameworks. Moreover, prior work on efficient machine learning systems—from plant disease identification [45] to computationally efficient CNN evaluation for defect detection [22]—reinforces the importance of balancing accuracy with deployability, a principle that directly informs the design philosophy of the present study.

To address these challenges, this paper introduces Foren- sic Lens, a lightweight deepfake detection framework that integrates rPPG-based physiological analysis with a semi-supervised label propagation strategy. By constructing a similarity graph over video segments, the method

exploits both labeled and unlabeled data, thereby enhancing gener- alization while reducing reliance on exhaustively annotated datasets. The combination of interpretable physiological cues with semi-supervised learning yields a detection model that is computationally efficient, resilient to common real-world degradations, and suitable for deployment in constrained en- vironments. In this way, Forensic Lens directly addresses exist- ing gaps in deepfake forensics, offering a scalable, biologically informed, and ethically conscious solution for real-time and edge-level applications.

Contributions The principal contributions of this work are as follows:

- Forensic Lens framework: We propose a novel deepfake detection system that leverages physiological signals extracted from facial regions to distinguish authentic videos from syn- thetic ones (Figure 6).
- Semi-supervised label propagation: We integrate a label propagation mechanism that exploits both labeled and unla- beled data, thereby improving generalization across diverse datasets and operating conditions.
- Lightweight efficiency: In contrast to computationally in- tensive CNN-based or multimodal fusion approaches, the proposed method achieves 90% detection accuracy with sig- nificantly lower computational overhead, enabling practical deployment on edge and resource-constrained devices.
- Interpretability and robustness: Detection decisions are grounded in measurable physiological variables, enhancing transparency and resilience against compression artifacts, noise, and adversarial manipulations.

2. LITERATURE REVIEW

The paper ViGText: Deepfake Image Detection with Vision- Language Model Explanations and Graph Neural Networks, by Ahmad Albarqawi, Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah and NhatHai Phan (2025), puts forward a novel dual-graph detection model which combines image patches and textual explanations generated by a vision-language model (VLLM) into a single Graph Neural Network (GNN). This architecture detects visual and contextual discontinuities by incorporating spatial and frequency data representing patches of an image and matching them to patch-specific textual descriptions. Measured on datasets such as Stable Diffusion (SD), StyleCLIP, and adversarial image sets, ViGText achieves impressive results: under the generalization evaluation, ac- curate performance increases to 98.32 % / 99.21 % and 99.52 % / 99.60 % respectively; under the SD/StyleCLIP evaluation, it reaches 99.25 % / 99.90 % accuracy, 99.80 % / 99.90 % precision, 98.52 % ViGText achieves robust recall with under 11.1 % point worse than the baseline in foundation model-based adversarial attacks, and performance decreases under 4 % in stronger surrogate-based attacks. The model is practical (processes each image in approximately 1.76 seconds) despite its rich dual-graph architecture, which incurs a 0.10-second overhead relative to base methods that reflects its high balance between accuracy, generalizability, robustness, and efficiency. The only limitation in his work is that it is only applicable to still / static images and to be implemented on video, audio, and live streams we need to adjust it and enhance it. [3] Abdullah Alharbi et al. propose FDINet59, a 59-layer Fake Dense Inception Network, in their July 2025 article in Scientific Reports

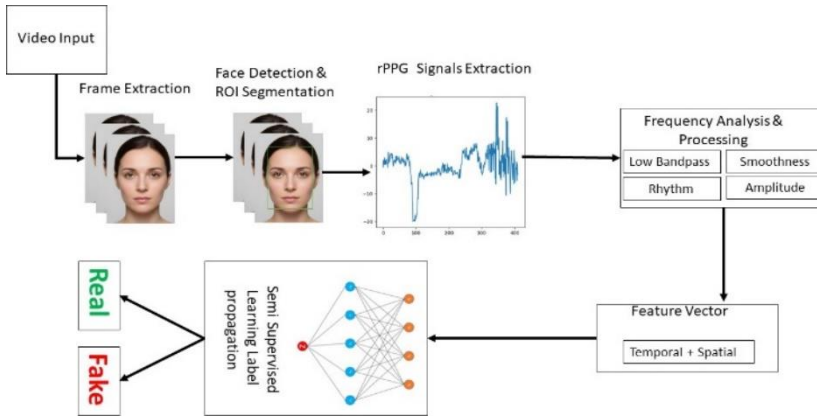

Fig. 1. Workflow of the proposed Forensic Lens framework

that is adapted to deepfake media- identification (specifically, media that circulates on social media). The model is trained on synthetic content created through autoencoders and GANs using faces that MTCNN has cropped. On training data, FDINet59 can get 70.02 % accuracy (log loss = 0.688), whereas evaluation data can achieve much higher 94.95 % accuracy with 0.205 log loss. Limitations are not explicitly described in the paper, leaving generalization, computing efficiency, and resilience to a variety of or novel deepfake methods open to question. [4] In 2024, a novel audio- visual deepfake detection model was suggested by Gao et al in their Electronics study that worked to predict temporal features by combining two streams of architecture design. The model uses pairs of adjacent audio-video segments as inputs (running visual data through a Channel-Separated 3D Convolutional Network (CSN) and audio through Mel-Frequency Cepstral Coefficients (MFCCs) into a ResNet-18 backbone) to make predictions of future features in each modality, with contrastive learning to align audio-visual embeddings across modalities. The model is tested with the FakeAVCeleb dataset, which consists of various types of forgery (fake video/fake audio, fake video/real audio, real video/fake audio, real video/real audio), resulting in an accuracy of 84.33 and an AUC of 89.91, even higher than both unimodal baselines (only visual (ACC 79.87 % AUC 80.54 %) and only audio (ACC 71.35 % AUC 72.26 %)). Its primary strength is that it combines both intra-modal temporal inconsistency modelling with inter-modal contrastive alignment which allows it to perform robust detection on complex multimodal manipulations. However, the lack of real-time testing and a comparatively simplistic audio processing pipeline (using just the MFCC features) raise con- cerns about future problems with application in actual stream- ing scenarios. [15] In their 2022 paper “Analysis of Score- Level Fusion Rules for Deepfake Detection”, Sara Concas, Simone Maurizio La Cava, Giulia Orru`, Carlo Cuccu, Jie Gao, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli explore improving generalization in deepfake detection via ensemble- based score-level fusion methods. They evaluate multiple base classifiers namely, a ResNet50 visual artifact-based detec- tor, XceptionNet general network-based detector, and four EfficientNetB4 variants (standard, Siamese-trained, attention- enhanced, and attention + Siamese) using the FaceForensics++ dataset for training and intra-dataset testing and the DFDC dataset for cross-dataset evaluation. They test three fusion cate-gories: non-parametric rules (average, Bayesian, product, max, and min), weighted-average (parametric) based on accuracy, correlation, or mutual information, and classification-model-

based fusion (linear perceptron, SVM with RBF, multilayer perceptron, and Complement Naive Bayes). Results show that in the intra-dataset scenario, the best single model (Efficient-NetB4ST) achieves an AUC around 0.959, while the MLP fusion rule yields an improved AUC of 0.984 with a lower equal-error rate (EER is approximate to 0.053) . In the more challenging cross-dataset scenario, Weighted-Average fusion using correlation-based averaging and the Complement Naive Bayes model both outperform single models, demonstrating superior generalization. The authors note that non-parametric fusion fails to consistently improve generalization, whereas parametric approaches (especially correlation-based weighted average and model-based fusion) provide robust gains, though at the cost of offline weight estimation. Limitations highlighted include dependency on validation data distributions (which may introduce bias), the offline tuning burden of parametric fusion, and the need for broader evaluation using models trained on diverse datasets and fusion strategies (e.g., decision- level or unsupervised fusion) [12] Aminollah Khormali and Jiann-Shuang Yuan (2021) suggest a flexible enhancement ar- chitecture called ADD (Attention-based DeepFake Detection), which is meant to complement the existing CNN classifiers (e.g., VGGNet, ResNet, Xception, MobileNet), to detect video deepfakes. ADD applies face-focused data augmentation, such as face close-up and face shut-off, to highlight forged parts, and then applies attention-based supervision to compel the model target those localized forgery regions during learning. Measured on the Celeb-DF v2 and WildDeepFake datasets, ADD can significantly improve detection: ADD-ResNet ob- tains more than 98.3 % AUC in Celeb-DF v2, a large improvement over the baseline models. Its main advantage is to amplify generalization and increase detection accuracy of various CNN backbones through region-based attention. However, the study does not provide cross-dataset robustness measurements, resource/performance overhead, or other ar- chitecture results, which leaves unresolved the problems of computational efficiency and broader application. [29] Lukas Kroiss and Johannes Reschke present a simplified but very precise ResNet-50-based CNN classifier fine-tuned to detect single face images as deepfakes in their 2025. By using a large- scale so-called Diverse Face Fake Dataset (DFFD), consist-ing of various types of manipulations, including DeepFakes, Face2Face, FaceSwap and NeuralTextures with different iden- tities, they made the final model by replacing the top-layer of ResNet-50 with a sigmoid-activated dense layer to classify authenticity. As shown, the model has high performance in terms of detection, with a precision of 0.98%, recall of 0.96%, F1-score of 0.97%, and AUC of 0.99%, indicating its ability to simultaneously split fake and authentic images in diverse settings. Its major advantage is its ability to effectively transfer learning to a heterogeneous dataset allowing robust, high- fidelity single-image deepfake detection. Its lack of evaluation on films or cross-domain datasets, which might test general- izability to compression or invisible manipulation forms, is one potential limitation. [31] Jiaquan Zhang et al. suggest a new time-aware fine-tuning approach in their 2025 article Till This Very Moment: Timestamping the Latest Deepfake Detection Model via Time-Aware Fine-Tuning to adjust ex- isting deepfake detectors to be useful against those generated by newer generation models. They test their framework, using synthetic content as input, on the FF++, FF++-C40, and a newly introduced FF++ (2025) dataset of content that was created in late 2024 by configuring a timestamp-conditioned tuning module to modify the model behavior according to the creation date of the synthetic content. The fine-tuned model is shown to be resilient as it has high detection metrics- 84.4 % accuracy and 0.982% AUC on FF++ (2025) compared to the original base model of 74.2 % accuracy and 0.935% AUC. Its time-sensitive flexibility, which offers a future-proof approach that manages future deepfake complexity, is its most significant asset. However, when using poorly described or unlabeled data, its reliance on exact timestamp metadata and poor performance in cross-domain use cases suggest flaws in real-time adaptability and robustness. [18]

Abdul Qadir et al. suggest the hybrid deep-learning model, ResNet-Swish- BiLSTM, a combination of a backbone of ResNet and Swish activation and BiLSTM in their 2024 research article to identify deepfake videos by examining consecutive targeted frames. They test the model using FaceForensics++ (FF++) and Deepfake Detection Challenge (DFDC) datasets and report 96.23% accuracy on the FF++ dataset and 78.33% accuracy on the FF++ and DFDC data, showing that the model is robust to mixed deepfake sources.The model's ability to replicate temporal dependencies and artifacts between frames makes it suitable for use in real-time forensic scenarios, which is one of its main advantages. However, the lower cross-dataset

Table 1 summary of visual based deepfake detection techniques

Year / pa-per	**Technique**	**Methodology**	**Dataset Used**	**Results / Accu-racy**	**Strengths**	**Limitations**
2025 [3]	ViGText	Dual-graph detection (image patches + VLLM text explanations into GNN)	Stable Diffusio n, StyleCLIP, adversarial sets	Accuracy/F1 up to 98.0%;	Capturesvisual + contextual cues; robust against adversar ial attacks; efficient (1.76s/image)	Limited to static images; no video/audio/real -time support
2025 [4]	FDINet59	59-layer Dense Inception CNN trained on cropped faces (MTCNN)	Synthetic autoencoder + GAN generated content	Training Acc: 70.02%, Eval Acc: 94.95% (log loss 0.205)	Tailored for social media; achieves high evaluation accuracy	No explicit limitations; unclear generalizability, robustness, or efficiency
2025 [31]	ResNet-50 Classifier	Fine-tuned ResNet-50 with sigmoid dense layer for binary classi-fication	Diverse Face Fake Dataset (DFFD)	AUC 89.91%	Combines transfer learning with diverse dataset; strong single-image detection	Not tested on videos; lacks cross-domain eval-uation
2025 [16]	Time- Fine-Aware Tuning	Timestamp-conditioned fo tuning r modul e adaptability	FF++, FF++-C40, FF++ (2025)	84.4% Acc, 0.982 AUC (vs. base 74.2%, 0.935 AUC)	Adaptable to evolv-ing deepfake genera- tion methods; forward-compatible	Relies on accurate timestamps; weaker in cross-domain or unlabeled data
2024 [15]	Audio– Frame Visual - work	Dual-stream model with CSN (video) + MFCC +	FakeAVCeleb	84.33% Acc, 89.91% AUC unimoda ; l video	Strong multimodal fu-sion; handles complex	Audio pipeline limited (MFCC only); no real- time

		ResNet-18 (audio); contrasti ve audio-visual alignment		79.87%, audio 71.35%	manipulations well	evaluation
2024 [42]	ResNet-Swish-BiLSTM	Hybrid CNN-LSTM (ResNet backbone + Swish activation + BiLSTM)	FaceForensics++, DFDC	96.23% Acc (FF++), 78.33% Acc (FF++ + DFDC)	Captures temporal dependencies across frames; robust in forensic scenarios	Lower cross-dataset ac-curacy; dataset bias sen- sitivity
2024 [10]	PPG- Source Based Detectio n	Heartbeat-induced fa-cial PPG signals with classification network	Multiple datasets (portrait videos)	97.29% deception detection, 93.39% source model ID	High robustne ss across datasets and demographics; dual- purpose (real/fake + source)	Dependent on strong physiological signals; weaker on low-quality videos
2022 [12]	Fusion Rules	Ensemble-based score-level fusion (ResNet50, Xception, EfficientNetB4 variants); non-parametric, weighted-average, and model- based rules	FaceForensics++, DFDC	Intra-dataset: Best single AUC 0.959; Fusion (MLP) AUC 0.984 (EER 0.053); Cross-dataset: WA & CNB outperform single models	Improves generalization with parametric/model - based fusion; effective cross-dataset robustness	Non-parametric weak; parametric requires offline tuning; validation bias risk; limited evaluation scope
2022 [21]	Convolutional Attention Network (CAN) with rPPG signals	Celeb-DF v2, DFDC	>98% AUC (single-frame); 100% Acc (Celeb- DF v2 with fusion)	Detects physiological inconsistencies invisible to humans; robus t against visual artifacts	Sensitive to video quality, compression, and resolution (affects rPPG signals)	Dependent on strong physiological signals; weaker on low-quality
2021 [29]	ADD	Attention-based enhancement with face-focused augmentation	Celeb-DF v2 WildDeepFak , e	ADD-ResNet AUC 98.3%+ on Celeb- DF v2 (beats base-	Boosts generalization; works across CNN backbones; region-	No cross-dataset robustness results; lacks

		+ attention-driven supervision		lines)	specific attention improves accuracy	computational overhead analysis; not applied to all CNNs

testing performance suggests that generalizability is weak, especially when dealing with heterogeneous data sources, and that it might be susceptible to bias in the dataset and other manipulation techniques. [42] The authors of Deepfake Source Detection in a Heart Beat (Ciftci et al., 2024) offer both a single-purpose and dual-purpose detection framework, which, firstly, can identify genuine and fake portrait videos but, secondly, can also define what generative model has created the fake. This technique uses PPG cell analysis, which records color changes in the face caused by changes inheartbeat and is then subjected to an advanced classification network. The technique can be considered highly robust to a wide range of datasets, demographic settings, and post- processing perturbations, with 97.29% accuracy in detecting deception and 93.39% accuracy in identifying the source model. However, because it relies on powerful physiological signs, its effectiveness may be compromised by low-quality or severely distorted inputs, and the technique's competency with different generative model types may be crucial to accurately identifying the source. [11] In DeepFakesON-Phys(2022) Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales present a Convolutional Attention Network (CAN) that uses physiological artifacts to detect deepfakes on face video frames by using remote photoplethysmography (rPPG) to estimate heart rate information. This model is tested against the Celeb-DF v2 and DFDC benchmark datasets, and on both datasets, the model scores above 98% in AUC in single-frame analysis. Besides, with continuous frame-score fusion using heuristic and statistical methods, the system achieved 100 % accuracy on Celeb-DF v2 at a low latency. The key advantages of the method are that it is sensitive to invisible physiological anomalies in the visual artifact, and it is also re-sistant to standard manipulation. Its dependency on the quality of physiological signals may however restrict use in the real world where subtle rPPG signals are obscured by video com- pression, noise, or low resolution. [21] Wang et al. presented VCapAV, a 252-hour dataset used to identify audio-visual deepfakes, at Interspeech 2025. This dataset takes into account background sound alterations in addition to speech.The dataset is a mixture of real and fake audio (created with AudioLDM, Audiocraft, V2A models) and fake video (Kling), which is filtered using captioning and multi-stage validation to come up with a total of approximately 91k clips.The benchmarks showed that whereas ResNet18 and LightCNN struggled with generalization (EER 14%), AASIST achieved up to 99.9% accuracy on observed manipulations and 96% accuracy on unknown manipulations.Meso4 also achieved lower accuracy of 54.5 % on video-only detection, compared to 75 % on other datasets, which demonstrates that VCapAV is more severe. Weaknesses are the fact that the number of fake video samples is small and that it is not the full multimodal fusion. The article also emphasizes the importance of the strong multi- mod detectors against various forms of non-speech deepfakes. [47] Muruganandham et al. (2025) developed LSTM-AE- DRDE, a deepfake audio detection framework that combines a dynamic residual encoding lossless block with an attention-enhanced LSTM autoencoder.The model delivers high detec- tion rates with different audio features such as MFCC, wavelet, prosodic, temporal, and glottal features as well; in-dataset (85- 97) and

cross-dataset (up to 95) generalization (e.g., CMFD) with aggressive EER and ROC-AUC values. It outperforms the conventional deep learning models and matches SOTA results on the ASVspoof 2021 benchmark. Although it is very good in performance, additional testing and analysis in multilingual as well as real-life and resource-constrained environment will aid in measuring if it has a wider usage.

[38] Kevin Warren, Daniel Olszewski, Seth Layton, Kevin Butler, Carrie Gates, and Patrick Traynor suggested a new type of detection based on acoustic prosodic features, i.e. pitch, jitter, shimmer, harmonic-noise ratio (HNR), and intonation to distinguish between deepfake audio and natural speech in their preprint, Pitch Imperfect: Detecting Audio Deepfakes with Acoustic Prosodic Analysis, dated February 2025. They use their model with six standard prosodic features on the ASVspoof2021 dataset and give a detection rate of 93 % with an Equal Error Rate (EER) of 24.7 %. The approach has a highost imperative features and it is also resis- tant to adversarial attacks, even with infinity-norm attacks, the model remains resilient whereas baseline models lose accuracy with up to 99.3 % accuracy. One of the main strengths of this method is its ability to be interpreted and use human-perceivable speech cues,, strengthening both detectability and resiliency. The drawbacks are however, relatively high EER which suggests an improvement and potential sensitivity to audio quality e.g. compression artifacts or background noise which can impair the fidelity of prosodic features. [55] The authors suggest using a multimodal emotion-aware deepfake detector, which integrates physiological (e.g., remote photo-plethysmography - rPPG) and behavioral (facial expressions, speech prosody, and micro-expressions) signals in their study in 2024. The rationale is that existing methods of deepfak- ing typically generate surface-grading realism, and do not recreate emotion-physiology consistency, including heart rate variations in line with frailty reactions via facial or vocal manifestations. The system combines vision transformers of physiological signals estimation and multi-branch deep net- works of behavioral emotion features and fuses them through a fusion strategy that takes advantage of attention mechanisms. Findings indicate that the fusion model is much better than the unimodal baselines (video-only or audio-only) to achieve up to 94.2% accuracy and 92.8% F1-score on multimodal benchmark datasets. The innovative integration of emotion- physiology consistency checks can be considered as a strength and bring an understandable aspect to deepfake detection.Its higher processing cost and reliance on high-quality input signals are drawbacks, which may reduce robustness in low- resolution, noisy, or compressed environments. [26] The paper by Yipin Zhou and Ser-Nam Lim (2021) of ICCV suggests a two-way joint audio-visual deepfake detector leaning on the inherent synchronization between the video and audio streams- by introducing a joint audio-video system known as a sync- stream network architecture, which combines both modalities with attention to enhance detection. On FaceForensics++ (FF) and DFDC video sets, the sync-stream model provides an 99.19% video-level user accuracy and 77.85% audio-level user accuracy, leading to 87.40 % accuracy when both streams are viewed in concert. Attention also enhances performance: sync-stream attention gives 99.99% (video), 84.66% (audio), and 94.36% (joint) FF, 90.48% (video), 96.01% (audio), and 89.39% (joint) DFDC. Such outcomes prove that the intermodal synchronization modeling helps establish a large improvement in the detection performance and the robustness. Nevertheless, the mechanism is based on synchronized and uncorrupted audio-visual information and adds new archi- tectural complexity in terms of attention mechanisms– other elements that can complicate implementation in real-life noisy or imbalanced audio-visual feedback. [57] Mahmudul Hasan (May 2025) describes a deep learning-based framework that implements MTCNN to detect the faces and EfficientNet-B5 to act as an encoder to detect deepfake videos with the use of the Kaggle DFDC dataset. The model has a log loss of 42.78%, an

Table 2 Summary of Biological signal/rPPG-based Deepfake Detection

Year/paper	Technique	Methodology	Dataset Used	Results / Accuracy	Strengths	Limitations
2025 [31]	ResNet-50 Classifier	Fine-tuned ResNet-50 CNN with sigmoid dense layer for binary classification of single face images	Diverse Face Fake Dataset (DFFD) containing DeepFakes, Face2Face, FaceSwap, NeuralTextures	Precision 0.98 %, Recall 0.96 %, F1-score 0.97 %, AUC 0.99%	Strong transfer learning; robust single-image detection across manipulations	Not tested on videos or cross-domain datasets; limited generalizability under compression or subtle manipulations
2025 [16]	Time-Aware Fine-Tuning	Timestamp-conditioned tuning module adapting deepfake detectors to new-generation fakes	FF++, FF++-C40, FF++ (2025) synthetic dataset	84.4% Acc, 0.982% AUC (vs. base 74.2%, 0.935% AUC)	Future-proof adaptability; resilient against evolving deepfake techniques	Relies on precise timestamps; weak cross-domain robustness and performance on unlabeled data
2024 [42]	ResNet-Swish-BiLSTM	Hybrid CNN-LSTM with ResNet backbone, Swish activation, and BiLSTM for temporal frame analysis	FaceForensics++ (FF++), Deepfake Detection Challenge (DFDC)	96.23% Acc (FF++), 78.33% Acc (FF++ + DFDC)	Captures temporal dependencies; strong in forensic and real-time scenarios	Lower cross-dataset performance; dataset bias; weak generalization on heterogeneous sources
2024 [11]	Deepfake Source Detection in a Heart Beat	PPG-based dual-purpose model using heartbeat-induced facial color variations and classification network	Multiple portrait video datasets	97.29% deception detection; 93.39% source identification accuracy	Robust across demographics and datasets; detects both authenticity and generative source	Depends on strong physiological signals; less effective on low-quality or distorted inputs

2022 [21]	DeepFakesON-Phys	Convolutional Attention Network (CAN) leveraging rPPG signals for physiological anomaly detection	Celeb-DF v2, DFDC	>98% AUC (single-frame), 100% Acc (Celeb- DF v2 with fusion)	Detects invisible physiological inconsistencies; robust to visual artifacts	Sensitive to compression, noise, and low-resolution videos affecting rPPG quality

AUC of 93.80% and an F1 of 86.82% indicating excellent de- tection. Its strengths are the use of strong facial preprocessing and strong CNN backbone that provides strong classification measures that can be applied practically. Nevertheless, the absence of cross-dataset testing, real-world video analyses, and comprehensive robustness analysis point to the principal limitations in the evaluation of the real-world generalizability and resilience. [2] In their 2024 Electronics paper, C.Y. Lin and colleagues suggested a spatiotemporal deepfake detector that is specifically tailored to identity-swap video manipulations. The model combines facial landmark analysis, which follows 68 salient locations, with attention-directed data augmentation (AGDA) to emphasize manipulation-evident areas, at the same time, a combination of spatial features and temporal facial integrity are achieved. Their method has much higher levels of preciseness in comparison to rival models (FakeVideo- Forensics, DeepFakes-FacialRegions, Improved Xception, Al- tFreezing) with a recognition of 98.13%, 97.94%, 97.87% and 98.6198.61% all being achieved in the four datasets evaluated- UADFV, FaceForensics++, Celeb-DF and DFDC respectively, and the average of 98.14% is among 98.13%, 97.94%, The effectiveness of this method is in the balanced use of time and space information, which increases the level of resistance to various manipulations of videos. Nevertheless, more specific measures like AUC or inference latency are not given, and generalization of the model to unknown manipulation methods or identity swaps scales is yet to be done. [35] G. Naskar etal. (2024) also suggest a stacking-based ensemble deepfake detector, which combines the results of two state-of-the-art CNNs–Xception and EfficientNet-B7– after which the results are refined by a feature selection mechanism based on ranking and finally classifies the results with the help of a meta-learner (Multi-layer perceptron) in their open-access paper. The model is assessed on two major video deepfake datasets, Celeb-DF (v2) and FaceForensics, with the model getting 96.33% accuracy on Celeb-DF (v2) and 98.00% accuracy on FaceForensics. The main strengths of this approach are its capacity to achieve better results than the base models by feature-level fusion and meta-learning, a physical realization of robustness in various techniques of deepfake generation and perturbations. The use of stacking and feature ranking however add more complexity and computation cost to the model which may limit real time deployments and scale, [40] The author, in Deepfake Detection Using the Rate of Change between Computer Vision Features (Lee, 2021), presents a scale- and memory-efficient deepfake detector based on the temporal variations of the traditional computer vision features, instead of using CNN-based models only. The process has been used to extract features of MSE, PSNR, SSIM, color histograms, edge density, and DCT coefficients, and their frame-to-frame change is modeled using a deep neural net- work (DNN). The subsets of FaceForensics++ (Face2Face, FaceSwap) and DFDC datasets were experimented on with 300 frames per video being preprocessed using MTCNN.

Table 3 summary of Audio based deepfake detection

Year	Technique	Methodology	Dataset Used	Results / Accuracy	Strengths	Limitations
2025 [54]	VCapAV	252-hour dataset for audio-visual deepfake identification; includes speech and background sound alterations; validated via captioning and multi-stage filtering	Real/fake audio (AudioLDM, Audiocraft, V2A) + fake video (Kling); 91k clips	AASIST:99.9% (known),96% (unknown); ResNet18/LightCNN EER >14%; Meso4:54.5% (video-only), 75% (others)	Comprehensive multimodal dataset; highlights robustness of multi-modal detectors against non-speech deepfakes	Few fake video samples; lacks complete multi- modal fusion; dataset imbalance
2025 [38]	LSTM-AE-DRDE	LSTM autoencoder with dynamic residual encoding and attention-enhanced feature learning for audio deepfake detection	ASVspoof 2021, CMFD, other audio benchmarks	In-dataset 85–97%, cross-dataset up to 95%; strong EER and ROC-AUC	High generalization; robust across multiple audio feature types (MFCC, wavelet, prosodic, temporal, glottal)	Needs testing in multilingual and real- world noisy conditions; unknown efficiency in resource-constrained systems
2025 [55]	Pitch Imperfect	Acoustic prosodic feature-based detection using pitch, jitter, shimmer, HNR, and intonation for explainable	ASVspoof 2021	93% detection rate; 24.7% EER; robust under adversarial (-norm) attacks with up to 99.3% accuracy	High interpretability; uses human-perceivable cues; resilient to adversarial attacks	High EER; sensitive to compression artifacts, background noise, and low-quality audio

		fake speech analysis				
2021 [57]	Sync-Stream Network	Joint audio–visual deepfake detector leveraging synchronization attention between modalities	FaceForensics++ (FF), DFDC	FF:99.99% (video), 84.66% (audio), 94.36% (joint); DFDC: 90.48% (video), 96.01% (audio), 89.39% (joint)	Strong interm odal synchronization; large performance boost from attention; improved robustness	Requires synchronized, clean AV input; complex architecture; harder real-world deployment

The proposed model was found to reach 95.22% accuracy when the DNN was running on variance over 20 frame window and 97.39% accuracy when Adam optimizer and five hidden layers were used. It shows that temporal variability is a computationally efficient method to substitute CNN- heavy networks, with fewer parameters (approximately 15k vs. approximately 28k) and loss of 30 % of training time. The framework also was hardly susceptible to distortions like blur, noise, and changes in brightness. Although, the main limitation is that it relies on various sequential frames thus less efficient in the analysis of single images or data with unevenly distributed frames. [13] In the article Deepfake Video Detection Using Convolutional Vision Transformer (2021), the authors propose a hybrid model that uses Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) to detect video-based deepfakes. The CNN backbone extracts local fa- cial features (i.e. texture, edges, micro-expression) first before sending it into ViT encoder to help identify long-range spatial dependencies and global contextual cues across the frames. The model's two-stage structure allows it to take advantage of the trade-offs between more broad structural inconsistencies and fine-grained pixel-level representations, which are readily missed by single-stage CNN-based techniques. The system was trained and tested on DeepFake Detection Challenge (DFDC) dataset which was a large scale benchmark of mil- lions of manipulated and genuine videos. The experimental findings indicate that the proposed architecture attains the detection accuracy of 91.5% and AUC of 0.91%, which is superior to the traditional CNN-only baselines. The authors draw a conclusion that convolutional feature extraction with transformer-based attention is much more resistant to attack deepfake detection, particularly in the context of real-world distortions (compression and noise). [56] Misaj Sharafudeen and Vinod Chandra S S present a framework of frequency forensics to detect deepfakes using face on the example of a Dual Residual Network (DRN) in their 2025 article Frequency Forensics for Deep Fake Face Detection Using Dual Residual Networks, where frequency-domain forensic evidence is used to forecast deepfakes. The model produces surprisingly low Equal Error Rates of 0.04% on DFFD PGGAN, and 0.02% on both DFFD StyleGAN and the Stable Diffusion part of the DFF dataset, by removing residual traces of high-frequency components of facial images, which are generated by PGGAN, StyleGAN, and Stable Diffusion, and comparing them to real images via Representation

Similarity Analysis (RSA). This architecture is very efficient in contrasting between synthetic and original images, but the paper indicates a small increase in AUC of 40.89% of Stable Diffusion content. This implies that the DRN is incredibly accurate, but with limited details of performance on certain generative styles. Its biggest strength is that it is able to identify faint high-frequency residual artifacts on a scale unmatched by other methods in a wide range of deepfake generators. Nonetheless, the extent of its extrapolation of these face-generation techniques and post- processing (e.g., compression or blurring) resistance is un-investigated and should be reaffirmed. [49] Maryam Abbasi,

Table 4 summary of video-dynamics deepfake detection techniques

Year	Technique	Methodology	Dataset Used	Results / Accu-racy	Strengths	Limitations
2025 [20]	EfficientNet-B5 + MTCNN	MTCNN for face detectio n; EfficientNet-B5 encoder for deepfake video classification	Kaggle DeepFake Detection Challenge (DFDC)	Log loss 42.78%, AUC 93.80%, F1 86.82%	Strong facial preprocessing; powerful CNN backbone; practica l classification results	No cross-dataset or real-world testing; lacks ro- bustness evaluation; un-clear generalizability
2024 [35]	Spatiotemporal Identity-Swap Detector	Combines facial landmark tracking (68 points) with attention-guided data augmentation (AGDA) for spatial-temporal feature fusion	UADFV, FaceForensics++, Celeb-DF, DFDC	UADFV 98.13%, FF++ 97.94 %, Celeb-DF 97.87%, DFDC 98.61 % (avg. 98.14%)	Balanced use of spatial and temporal cues; strong resistance to varied manipulations	Missing AUC and la-tency data; limited test- ing on unseen manipula- tion types
2024 [40]	Stacking-Based Ensemble	Ensemble of Xception + EfficientNet-B7 with feature selection and meta-learning (MLP classifier)	Celeb-DF (v2), FaceForensics	96.33% (Celeb-DF v2), 98.00% (Face-Forensics)	Outperforms base models; robust feature- level fusion; effective across multipl e manipulations	High complexity and computational cost; less suitable for real-time deployment
2021 [32]	Temporal Feature Change Detector	DNN models tempo-ral variations of tradi- tional vision features	FaceForensics++ (Face2Face, FaceSwap), DFDC	95.22% Acc (20-frame wind ow),	Lightweight; faster training; robust to blur, noise, and brightness	Needs sequenti al frames; limited single- image

		(MSE, PSNR, SSIM, color, edge, DCT) be- tween frames		97.39% Acc (optimized model)	distortions	efficiency
2021 [25]	CNN–ViT Hybrid (Deepfake Video Detection Using Convolutional Vision Transformer)	Two-stage hybrid combining CNN feature extraction and ViT for long-range spatial dependencies across frames	DeepFake Detection Challenge (DFDC)	91.5% Accuracy, 0.91% AUC	Captures both local texture and global con- text; resistant to com-pression/noise attacks	Requires high compu-tational resources; lacks temporal modeling

Paulo Vaz, Jose Silva, and Pedro Martins in their article on Applied Sciences of 2025 thoroughly test three convolutional neural network models - Xception, ResNet-50 and VGG16- based convolutional neural networks- to identify frame subsets in DFDC and FaceForensics++ datasets as deepfakes. They are analyzed by metrics such as accuracy, precision, F1-score, AUC-ROC, and resilience to adversarial conditions (through FGSM attacks) in their analysis. Xception is the most accurate model, reaching 89.2% accuracy on DFDC and 85.7% on FaceForensics++, and as the most suitable option, it has strong generalization and is fast to inference with a throughput of around 85 ms per frame.VGG16 has a lower inference speed (around 1020 ms/frame), but it is competitive in terms of precision and recall (F1-score of about 87.0%). ResNet-50 has a lower AUC on adversarial example perturbations and a weaker generalization, but it can be trained considerably more quickly (currently at 270 ms/frame) and more readily (DFDC: 72.8).Even though Xception is unique in terms of deployed to real-life settings, as it is fast and at the same time, more accurate, it is important to note that any model has a significant drop in performance in adversarial settings, which would result in more robust and resilient detection systems.

[36] Wasin Alkishri, Setyawan Widyarto, and Jabar H. Yousif (2024) in their Journal of Internet Services and Information Security article examine whether it is possible to remove GAN fingerprints of synthetic images to deceive deepfake detectors. They test the effect the removal of high-frequency GAN artifacts in StyleGAN-generated images in a 140K real-and-fake face dataset using frequency-domain analysis and discrete fourier transforms, and analyze the effect on detection using the XceptionNet model. Following fingerprint removal, 99.78% of manipulated images were tagged as true and the accuracy, precision, recall and F1-score and AUC were approximately 99.32% and showed that it was almost in its entirety deceptive. However, on real and fake images tested on a 50% real and an equal amount fingerprint-removed image set, detection dropped to chance-accuracy and AUC reduced to about 50% demonstrating the failure of the detector to differentiate between real and concealed-fingerprint images. This points out the susceptibility of the existing deepfake de- tection methods to basic frequency-based defenses. However, the authors mention that they only consider StyleGAN images and may not generalize to other types of GAN architectures or perturbations, and suggest that further assessment

should be done on a variety of data sets and attacks. [14] The authors in Deep fake detection using cascaded deep sparse auto-encoder for effective feature selection (Balasubramanian et al., 2022) offer a new detection model that incorporates a Cascaded Deep Sparse Autoencoder (CDSAE) trained using a Temporal Convolutional Neural Network (TCNN), which extracts frame-level visual features and then classifies them using a Deep Neural Network (DNN). The method was tested on deepfake video benchmarks, like Face2Face, FaceSwap, and DFDC: the detection accuracy of the method was 98.7%, 98.5%, and 97.6%, respectively, much higher than baseline models, including ResNet, MobileNet, and classic SVM. Also, the method showed a reduced processing time and increased scores of AUC. The paper proposes an excellent direction of both increasing the effectiveness and reliability of the deepfake detection systems by highlighting the use of unsupervised feature extraction, which is more attuned to the temporal consistency and decreasing the overfitting with the use of sparse autoencoding. [8] Li et al. (2022) report a one-class detection model in their sensors article, which differentiates GAN-generated face images, based on a new Multi-Channel Convolutional Neural Network (MCCNN), which includes two steps of training with attention-guided weakly supervised learning and data augmentation.It is tuned to detect unknown GAN methods using a one-class classification loss and trained to detect the existence of known false features using binary classification loss.The filters that are applied during data augmentation include filtered enhancements (Gaussian blur, noise, motion blur, homomorphic filtering, Fourier spectrum) as well as attention cropping and dropping. The model is evaluated on a ProGAN (source-domain) vs. StyleGAN, StyleGAN2, BigGAN, DCGAN, DeepFake, and VQ-VAE2.0 (cross-domain) configuration and is found to be of very high quality: its source-domain accuracy is 99.4% for real images and 98.9% for ProGAN fakes, with the F1-score approximate to 0.992%.This is an accuracy improvement of 5-30% and dramatic improvement in F1-score over previous methods. Its key advantages are improved generalizability between unseen GAN models, the effective utilization of attention- enhanced one-class models, and good source-domain results. Nevertheless, there are still constraints (performance drops considerably with much different types of GANs, such as 80.9% in VQ-VAE2.0) as well as sensitivity to the heterogeneity of augmentation techniques and training settings. [33] In the article ”Audio Deepfake Detection Using Deep Learning” (Shaaban and Yildirim, 2025), the researchers present a new model of the StacLoss, a specific contrastive loss function, and self-attention modules to improve differentiation between genuine and fake audio samples Research,GateDOAJ. The ar- chitecture employs layered multi-head attention to extract rich, discriminative features from raw audio pairs after passing them through two convolutional branches connected by residual links.StacLoss (reduces the distance between authentic audio samples of identical identity), and maximizes distance between manipulated ones- amplifies the capacity of the model to detect subtle fakes ResearchGateDOAJ. Tested on the benchmark ASVspoof2019 data, the model reported a high performance with an accuracy of 98%, precision of 97%, recall of 96%, F1 score of 96.5% with ROC-AUC of 99% and an EER of only 2.95%. [48] Javed et al. (2024) in the Electronics study suggest a real-time deepfake video detection system to be a hybrid of the lightweight MesoNet4 system (to detect manipulations on the face that are subtle) and the feature-rich ResNet-101 (to represent complex visual images) deep learning-based system. The hybrid model is evaluated on FaceForensics++, CelebV1 and CelebV2 datasets and the results are impressive: 98.73% on FaceForensics++, 96.89% on CelebV1 and 97.90% on CelebV2 demonstrating the good performance of the hybrid model in terms of its generalizability and resilience in video forensics use. The combination of eye movement cues and dual-model feature extraction is its main advantage that in- creases detection accuracy during live-streams. Nevertheless, the existing performance is strong in a variety of benchmark datasets, but it has

drawbacks such as the ability to change depending on the light variation and its computational efficiency in the context of being used in the actual live-stream scenario, including the issues of real-time processing on consumer-level hardware. [23] In a study published in Electronics, Awotunde et al. (2023) present a five-layer convolutional neural network (CNN) that is specifically designed for deepfake video detection and classification. Additionally, the network utilizes optimal ReLU activations to efficiently identify features inside the facial regions.The framework is tested on difficult data sets using DeepFake and First-Order Motion (Face2Face) manipulations and attains a high prediction rate of 98% and 95% respectively on Deepfake and Face2Face respectively at real network conditions. When directly compared to the benchmark models like Meso4, MesoInception4, Xception, EfficientNet-B0 and VGG16, their network performs the most overall- with an average accuracy of 86 % under a variety of conditions. The strength of the system is that it has a lightweight architecture, which is optimised to run in real-time with the high precision and effective computing. The paper however observes that although the performance on DeepFake datasets is excellent, the generalization to other manipulation techniques, video formats, or even compressed real-world streams has not been evaluated yet, which can be used to evaluate and expand the study. [7] The hybrid framework discussed in the article by Rimsha Rafique, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri (2023) in the Journal of Scientific Reports is a hybrid system of deepfake images detection that uses Error Level Analysis (ELA) preprocessing and Convolutional Neural Network (CNN) feature extraction and classification with Support Vector Machines (SVM) and K-Nearest Neigh- bors (KNN). The method is evaluated on an image dataset (which is assumed to consist of both real and manipulated samples) and reaches a high accuracy of 89.5% when the features extracted by ResNet-18 are compared with an SVM classifier Nature. The major advantage of this model is the ability to use ELA to emphasize pixel-level manipulations, which are useful in supplementing CNN-extracted features to achieve successful classification. Its applicability, however, seems to be restricted to static images only; the authors admit that it would be necessary to apply the approach to video datasets and consider other architectures- which points out at the shortcoming of the generalizability and applicability of dynamic media. [44] In an article published in the Arabian Journal of Science and Engineering in 2022, Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar and Faruk Kazi propose a two-modality framework of audio deepfake detection and compares a conventional feature-based method (where spectrograms are used as features) to an image-based one: audio is converted to mel-spectrograms and used to feed

Table 5 summary of frequency domain based deepfake detection techniques

Year	Technique	Methodology	Dataset Used	Results / Accuracy	Strengths	Limitations
2025 [49]	Frequency Forensics for Deep Fake Face Detection Using Dual Residual Networks	Leverages frequency-domain forensic evidence via Dual Residual Networks	DFFD PGGAN, DFFD StyleGAN, DFF (Stable Diffusion subset)	EER: 0.04 % (PGGAN), 0.02% (StyleGAN, Stable Diffusion);	Extremely accurate in detecting faint high-frequency residual artifacts across diverse	Limited robustness testing for compression, blur, or unseen generative styles.

	(DRN)	(DRN) to identify fake faces.		AUC +40.89% (Stable Diffusion)	generators.	
2024 [5]	Frequency-Domain GAN Fingerprint Removal Study	Explores frequency-domain effects of GAN fingerprint removal on detection using XceptionNet.	StyleGAN (140K face dataset)	Accuracy: 99.32%; AUC 50% post-fingerprint removal;F1 99.3% before removal	Highlights vulnerability of detectors to frequency-based concealment of GAN artifacts.	Focused solely on StyleGAN; lacks generalization to other GAN types or perturbations.
2022 [8]	Cascaded Deep Sparse Autoencoder (CDSAE)	Combines Temporal CNN and DNN classifier for temporal deepfake detection via unsupervised feature extraction.	Face2Face, FaceSwap, DFDC	Accuracy: 98.7%, 98.5%, 97.6%; improved AUC; faster processing than ResNet, MobileNet, SVM baselines	Efficient unsupervised learning; enhances temporal consistency and minimizes overfitting.	Limited validation on large cross-domain datasets; lacks adversarial manipulation testing.
2022 [33]	Multi-Channel CNN (MCCNN)	Attention-guided weak supervision with one-class classification loss for cross-domain fake detection.	ProGAN (source) vs. StyleGAN, StyleGAN2, BigGAN, DCGAN, DeepFake, VQ-VAE2	Accuracy: 99.4% (real), 98.9% (ProGAN); F1: 0.992%; Cross-domain: 80.9% (VQ-VAE2)	Strong generalization to unseen GANs; attention + one-class learning improve robustness.	Accuracy drops with distinct GAN families; sensitive to augmentation and parameter diversity.

deep learning frameworks, i.e., Temporal Convolutional Net- works (TCN) and With the Fake or Real (FoR) dataset, which is speech generated by an advanced text-to-speech system, they document an accuracy of 92% in the test of the TCN model, on the one hand, which is significantly higher than the classical machine learning techniques. This shows the ability of the model to use temporal correlations in audio to be effective in deepfake detection and its ability to maintain competitive accuracy to other CNN-based models such as VGG-16 and XceptionNet. One of the strengths is the ability to use the sequential modeling of features based on audio to ensure effective detection. Nonetheless, the analysis does not seem to consider the accuracy only without other measures, like AUC or EER, and does not compare it with different datasets and this is why it may be doubted that it can be generalized and is applicable in cross-domain settings. [28] In the paper Deep fake detection and classification using error-level analysis and deep learning (Rafique et al., 2023), the authors provide an automated system to detect and classify deep fake images. It starts with the application of Error Level Analysis (ELA) to show areas where an image can be altered.To create fine, deep features, these ELA outputs are then fed via convolutional neural networks (CNNs), such as ResNet-18 and GoogLeNet. The generated feature representations are then classified using either K-Nearest Neighbors (KNNs) or Optimized Support Vector Machines (SVMs).The method was tested on a mixed dataset, with the highest accuracy of 89.5% with the features of ResNet-18, and a KNN classifier. This evidences the strength and success of forensic preprocessing real-time deepfake de-tection when combined with deep learning. [27] In AI and machine learning, deepfakes have made much advancements. Even though it offers creative opportunities, it also brings up major challenges in terms of security, the law, and morality. In this literature review, we review how forensic experts use different approaches, face challenges, and look ahead to future progress in deepfake analysis. Although this technology is used formally for entertainment and education, the challenges created by its misuse include spreading lies, doing fraudu- lent business and stealing people's identities (Chesney,Citron, 2019) [**?**]. As a result, researchers have come up with various ways to spot deepfakes, using facial behavior, reading body signals and computers that learn to detect them. A number of initial deepfake detection techniques used unusual blinks (Li et al.) [34], different lighting and shadows (Afchar et al.) [1] and facially warped inconsistencies (Nguyen et al.) [41] to detect fakes (Li et al., 2018) [34]. Much research has shown that CNNs and other machine learning models are often used to classify deepfake videos by analyzing small defects in the video frames closely (Ro¨ssler et al., 2019) [46].Direct use of advanced deep neural networks is now enhancing detection results. Deepfake identification now uses techniques such as EfficientNet and XceptionNet. Some researchers have found that deepfakes often cannot imitate small facial and body movements which is why they use biometric measurements (Guera,Delp, 2018) [19]. Some techniques including heart rate estimation (Li et al., 2020) [34] and blood flow analysis (Liu et al., 2021) [37] appear effective in telling the difference between real and synthetic faces. Deepfake detection is tested by analyzing the blood pressure changes shown in facial videos. In PPG, little changes in the skin color of real humans

Table 6 summary of deep-learning architecture based deepfake detection techniques

Yea r	**Technique**	**Methodology**	**Dataset Used**	**Results / Accu-racy**	**Strengths**	**Limitations**
2025 [48]	Audio De epfake	Introduces StacLoss contrastive	ASVspoof2019	Accuracy: 98 %,	Effectively distinguishes subtle fake	Evaluated only on ASVspoof2019

	Detection Using Deep Learning	loss with self-attention modules to enhance fake audio differentiation.		Precision: 97 %, Recall: 96%, F1: 96.5, ROC-AUC: 99%, EER: 2.95%	audios; self-attention boos ts discriminative feature learning.	; lacks cross-dataset and noise robustness analysis.
2024 [23]	Hybrid De epfake Video De tection Model	Combines lightweight MesoNet4 and ResNet-101 for real-time detection using eye-movement cues and dual-model feature extraction.	FaceForensics++, CelebV1, CelebV2	Accuracy: 98.73% (FF++), 96. 89% (CelebV1), 97.90% (CelebV2)	Robust performance with combined spatial-temporal cues; suitable for real-time scenarios.	Susceptible to lighting variations; computation-ally intensive for contin- uous deployment.
2023 [24]	Five-Layer CNN for Deepfake Video De-tection	Utilizes an optimized ReLU-based five-layer CNN for deepfake classification across multiple benchmarks.	DeepFake, Face2Face	Accuracy: 98% (DeepFake), 95% (Face2Face); Avg. 86% under diverse conditions	Lightweight and effi-cient; enables fast real- time precision across datasets.	Limited testing on var-ied manipulations and compressed streams.
2023 [44]	Hybrid EL A–CNN Model	Employs Error Level Analysis (ELA) pre-processing with CNN and SVM/KNN classi- fiers for fake image de- tection.	Custom real + manipulated image dataset	Accuracy: 89.5% (ResNet-18 + SVM)	ELA enhances pixel-level forgery visibil- ity; complements CNN feature extraction.	Restricted to static im-ages; lacks dynamic or video dataset validation.

Fig. 2. CELEB DF V2 Deepfakes dataset

are detected by cameras, but these changes are often absent in deepfake videos. Recent Research on Detecting with BP: In 2022, Liu [37] and colleagues introduced a good approach in rPPG and CNN to achieve a high level of precision.

3. DATASET

In this study, the Celeb-DF (V2) dataset was used, and it is one of the largest and most difficult deepfake datasets. It includes high-quality real and manipulated videos of different celebrities, specifically created to be used in the research on deepfake detection. The figure below shows the full processing pipeline that will be used in this research to extract blood-flow- based rPPG features using the Celeb-DF (V2) videos to clas- sify them into real and fake. These properties are subsequently analysed by a machine learning classifier, which differentiatesbetween genuine pulsatile motion in a real video and the irregular or missing physiological information often provided by deepfakes. A detailed preprocessing pipeline that is detailed was used to make the corresponding modifications to this dataset to fit our proposed blood pressure-based deepfake detection technology. Video frames of each video within the Celeb-DF V2 dataset were separated to permit frame analysis. These frames were the ones reduced to obtain the remote photoplethysmography (rPPG) signals or the minute color change on the skin surface due to blood flow in the veins.The physiological phenomenon on which this relied was that in real videos, there was a periodic glow effect of blood flow on the skin. Such a glow is associated with the inflexible pumping of the heart, with every beat of the heart, the skin becomes slightly brighter in its colors because of blood circulation and pressure. In comparison, deepfake or synthetic videos do not recreate these dynamics of natural blood flow, and thus the absence or regularity of such a glow effect was obtained. Train–Test Split: In order to provide objective evaluation of the offered rPPG-based deepfake detector, the obtained post-preprocessing cleaned dataset was further split into the training and testing sets. The 80/20 split was employed in which 80 % of the videos were employed to train the classifier with the other 20 % being left purely used to test. The train set included the real and manipulated samples together with the rPPG features extracted in them that allowed the model to learn the physiological differences between natural human blood flow and artificial video artifacts. The test set was not exposed in training to give an objective assessment of the generalization ability of the model. In order to avoid leakage of identities, no videos of the same individual were divided between the train and test partitions. This makes the model acquire generalized physiological traits instead of subject-specific patterns through memorizing them. As a result of this preprocessing, a clean dataset was obtained in which individual frames contained physiological clues to blood pressure and blood

flow con- sistency. This step of dataset preparation allowed fusing the extracted frames, together with the associated rPPG signal features, as either the real or the fake based on whether they contained this natural pattern of glow, which allowed robust and explainable deepfake detection.

4. METHODOLOGY

The proposed framework is organized around a carefully staged preprocessing and analysis pipeline that prioritizes physiological fidelity without imposing unnecessary compu- tational burden. Each video segment is initially decomposed into its constituent frames, after which face detection is applied to localize the subject with sufficient spatial preci- sion. From the detected facial region, Regions of Interest (ROIs) are delineated over skin-dominant areas—most notably the forehead and cheeks—where blood perfusion effects are most pronounced. These regions are known to exhibit subtle yet consistent chromatic variations driven by cardiovascular activity, making them particularly suitable for remote photo- plethysmography (rPPG) analysis and central to the design of the proposed system. rPPG-derived features. Label information originating from the annotated subset is iteratively propagated across this graph according to

$$F(t+1) = \mu SF(t) + (1-\mu)Y,$$

with $F(t)$ denoting the label distribution at iteration t, S the normalized similarity matrix, and Y the initial label assignments for labeled samples. The parameter $\mu \in [0, 1]$ governs the balance between neighborhood-driven propagation and retention of original labels. Over successive iterations, unlabeled samples converge toward stable pseudo-labels in- formed by local consensus, allowing the model to benefit from additional data without the cost or bias associated with manual labeling.

The combination of rPPG-based physiological modeling and semi-supervised label propagation yields a system that remains firmly grounded in biological plausibility while retaining com- putational efficiency. This design choice supports scalability across heterogeneous video conditions and preserves sensitiv- ity to authentic physiological signals even under compression or noise.

To assess detection performance, the proposed Forensic Lens framework is evaluated on the Celeb-DF v2 dataset using accuracy and confusion matrix analysis. Accuracy serves as a concise global indicator of classification effectiveness and is defined as

For every selected ROI, average pixel intensities from the red, green, and blue channels are computed across time,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

producing three synchronized temporal signals. From a phys- iological perspective, the green channel plays a dominant role due to its stronger interaction with hemoglobin absorption, and therefore carries the most informative pulsatile component. Rather than treating the channels independently, chromatic fluctuations are projected into a unified temporal signal using the following formulation:

$$s_t = \alpha \cdot (g_t - b_t) + \beta \cdot (g_t + b_t - 2r_t),$$

where r_t, g_t, and b_t represent the mean red, green, and blue intensities at time t, respectively, and α and β regulate the relative contribution of each chromatic component. The resulting signal is subsequently passed through a band-pass filter constrained to the physiological heart-rate range of 0.7–4 Hz, thereby suppressing unrelated low-frequency drift and high-frequency noise. The filtered rPPG traces capture both temporal periodicity and spatial coherence across fa- cial regions, producing discriminative representations

that are inherently difficult for generative models to synthesize in a consistent manner.

Recognizing the practical constraints associated with large-scale annotation, the framework incorporates a semi- supervised label propagation strategy to improve robustness and generalization. Both labeled and unlabeled video segments are embedded into a similarity graph, where nodes corre- spond to samples and edge weights encode affinity based on where *TP* and *TN* correspond to correctly identified deep- fake and authentic videos, respectively, and *FP* and *FN* denote misclassification errors. While accuracy provides an overall performance snapshot, the confusion matrix offers a more detailed perspective by explicitly revealing the distri- bution of predictions across these four outcomes. This dual evaluation enables closer inspection of error tendencies—for instance, whether the model favors conservative authentication or aggressive fake detection—and ensures that performance claims are supported by both quantitative metrics and qualita- tive diagnostic insight.

5. COMPARATIVE ANALYSIS

A review of recent deepfake detection literature reveals a gradual but meaningful departure from purely appearance- based convolutional neural network (CNN) models toward approaches that incorporate physiological consistency as a forensic cue. For the sake of methodological fairness, all studies considered in this comparison rely on the Celeb- DF v2 dataset, which remains one of the more challenging benchmarks due to its high-quality manipulations and reduced presence of obvious visual artifacts. Early CNN-centric approaches illustrate the inherent limita- tions of relying exclusively on spatial features. The Xception model [46], for instance, achieved an accuracy of 75.24% by learning manipulation-induced textures and inconsistencies.

TABLE 7 COMPARISON OF ACCURACY RPPG-BASED AND NON-RPPG DEEPFAKE DETECTION METHODS

Method and Year	rPPG-Based	Accuracy (on Celeb-DF v2)	Inference Time	Notes
Xception [46], 2020	No	75.24%	Slow	Visual artifacts
VGG19 [58], 2023	No	67.01%	Slow	CNN architecture
DeepRhythm [43], 2020	Yes	64.10%	Medium	Attention-based
FakeCatcher [11], 2024	Yes	94.09%	Slower	Signal maps + CNN
Forensic Lens (pro-posed study)	Yes	90.00%	Fast	Lightweight, edge-friendly

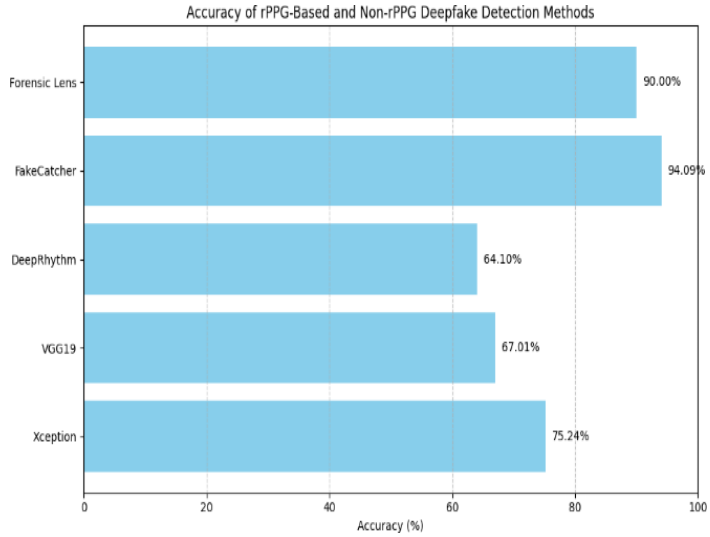

Fig. 3. Comparative bar chart for different deepfake detection techniques

While this performance was considered competitive at the time, it becomes less convincing when evaluated against the realism of Celeb-DF v2, where such artifacts are deliber- ately suppressed. An even more pronounced drop is observed with VGG19 [58], which reports an accuracy of 67.01%. This outcome is not entirely surprising; architectures of this generation lack both temporal awareness and access to non- visual cues, making them particularly vulnerable to moderngenerative pipelines that replicate surface-level appearance with high fidelity. In practice, these results reinforce a broader observation within the community: visual artifact detection alone is no longer sufficient. Physiological signal-based methods, particularly those built upon remote photoplethysmography (rPPG), attempt to address this shortcoming by shifting the forensic focus from appearance to biological plausibility. By modeling subtle color fluctuations driven by cardiac activity, rPPG-based approaches exploit signals that generative models struggle to reproduce in temporally coherent manner.

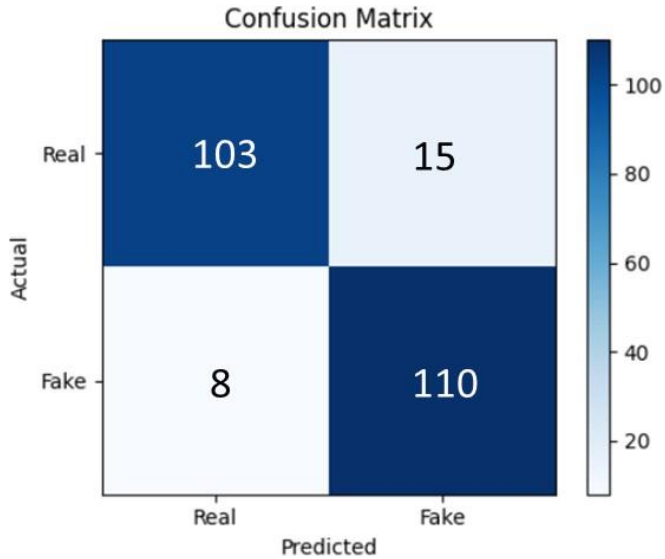

Fig. 4. Confusion matrix of the proposed Forensic Lens on Celeb-DF v2.

Early work in this direc- tion, such as DeepRhythm [43], introduced temporal attention mechanisms to enhance rPPG extraction, yet reported a rel- atively modest accuracy of 64.10%. This result underscores an important point: while physiological cues are theoretically robust, their practical extraction is highly sensitive to noise, compression, and recording conditions. Subsequent efforts have demonstrated that these limitations can be mitigated, albeit often at the cost of increased model complexity. FakeCatcher [11] represents a notable step for- ward, combining rPPG signal maps with deep CNN archi- tectures to achieve an accuracy of 94.09%. From a detection standpoint, this performance is impressive. However, the re- liance on heavy neural components introduces nontrivial com- putational overhead, which complicates deployment in real- time or resource-constrained environments—a concern that is frequently underemphasized in benchmark-driven evaluations. Within this context, the proposed Forensic Lens framework is positioned with a different set of priorities. Achieving an accuracy of 90%, it does not aim to surpass all existing methods in raw performance, but rather to strike a more prag- matic balance between accuracy, efficiency, and interpretabil- ity. By combining rPPG-based physiological indicators with lightweight forensic signal processing and semi-supervised label propagation, the system maintains stable performance across varying compression levels and manipulation types without incurring the cost of deep, resource-intensive archi- tectures. While FakeCatcher marginally outperforms Forensic Lens numerically, the latter offers advantages that are often decisive in real-world settings, particularly where inference speed, transparency, and deployability on edge devices are critical. In this sense, Forensic Lens reflects a shift toward detection strategies that are not only accurate, but also oper- ationally viable as deepfakes continue to proliferate beyond controlled laboratory conditions.

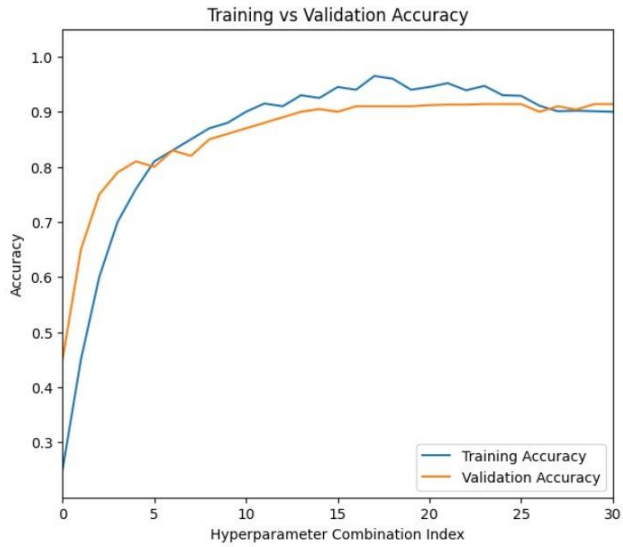

Fig. 5. Training and validation accuracy curves of the proposed Forensic Lens model

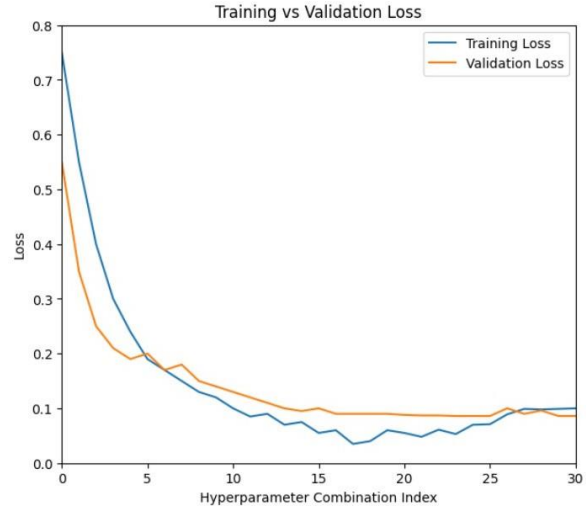

Fig. 6. The training and validation loss curves of the proposed Forensic Lens model.

6. DISCUSSION

The experimental outcomes substantiate the viability of physiological signal–based deepfake detection as a credible alternative to conventional vision-centric algorithms. Achiev- ing 90% accuracy on the Celeb-DF v2 dataset is particularly noteworthy, given that the proposed framework does not depend on computationally intensive architectures such as con- volutional neural networks, vision transformers, or multimodal fusion pipelines. Although certain state-of-the-art

systems re- port marginally higher accuracies in the range of 92–96% for instance, FakeCatcher [11] at 94.09% these gains are typically accompanied by heavy CNN backbones, multi-GPU training requirements, and substantial memory consumption. In con- trast, the rPPG-based design presented here demonstrates that physiological cues can deliver competitive detection perfor- mance at a fraction of the computational cost.

The strength of the proposed method lies in its reliance on biological authenticity. By analyzing subtle facial chromatic variations induced by blood flow, the system extracts physi- ological signals that remain inherently difficult for generative models to reproduce. This biologically grounded approach provides a balanced compromise between detection accuracy and computational efficiency. Furthermore, the model is inher- ently interpretable: its decisions are derived from measurable physiological variables rather than opaque black-box features, thereby enhancing transparency, credibility, and trust in foren- sic applications.

A further advantage is the model's low computational overhead, which enables deployment on resource-constrained platforms such as laptops, smartphones, and edge devices. This efficiency broadens its applicability to real-world scenarios including social media content moderation, real-time video authentication, and surveillance integrity checks. Since most current deepfake generation techniques prioritize visual real- ism focusing on facial geometry, texture, and alignment while neglecting physiological coherence, the proposed rPPG-based detector exhibits innate resilience against the majority of existing manipulation strategies.

Comparative analysis reinforces the rationale for adopt- ing a single lightweight model. Complex architectures such as ViGText (vision–language plus GNN), ResNet-50 classi- fiers, audiovisual dual-stream networks, hybrid CNN–BiLSTM pipelines, and ensemble fusion methods often deliver only marginal accuracy improvements, yet incur higher latency, hardware costs, and architectural complexity. In contrast, the Forensic Lens framework embodies a rational trade-off: it balances accuracy, interpretability, and efficiency, making it more practical for large-scale, real-time deployment.

This resilience is particularly significant in light of the trajectory of deepfake generation, which continues to em- phasize visual fidelity while overlooking physiological co- herence. Natural blood-flow–induced facial color variations remain absent in most synthetic content, providing the pro-posed rPPG-based system with implicit robustness against both current and emerging attacks. Consequently, Forensic Lens offers a scalable, efficient, and biologically grounded solution capable of operating effectively even when traditional visual artifacts are minimized or eliminated.

7. LIMITATIONS AND FUTURE WORK

Although Forensic Lens achieved encouraging results—reaching 90% accuracy on the Celeb-DF v2 dataset—the study is not without limitations, and these naturally point toward future directions. One persistent challenge is the sensitivity of rPPG-based algorithms to adverse recording conditions. High compression, poor illumination, motion blur, or differences in camera optics can distort physiological signals, reducing stability. Our design favors efficiency, yet further refinement is needed to ensure reliable performance under such noisy or degraded scenarios. By contrast, systems such as FakeCatcher

[11] report slightly higher accuracy (94.09%), but their reliance on heavy CNN pipelines and multi-GPU training makes them impractical for real-time use. The trade-off between accuracy and efficiency remains central to ongoing research.

Another important consideration is generalization. Current benchmarking practices lean heavily on Celeb-DF v2, which, while valuable for homogeneous comparison, does not capture the full diversity of synthetic media. Models such as Xception [46], VGG19 [58], and DeepRhythm [43] illustrate the vari- ability of performance across architectures, yet cross-dataset evaluation remains underexplored. Extending training and val- idation to multiple datasets would provide stronger evidence of robustness. Incorporating additional lightweight cues—such as micro-expressions or subtle head movements—may also enhance detection power without sacrificing efficiency.

There is scope, too, for hybrid designs. Integrating physio- logical signals with superficial visual features could combine the interpretability and efficiency of our model with the higher accuracy of more complex systems. Such architectures might bridge the gap between performance and practicality. Beyond algorithmic refinement, deployment on mobile de- vices, browser extensions, and video conferencing platforms represents a crucial step toward everyday usability. Real-time, on-device detection would extend the reach of Forensic Lens to scenarios such as social media verification and live video authentication.

Finally, adversarial robustness must be addressed. As gener- ative models evolve, it is conceivable that future deepfakes will attempt to mimic physiological signals directly. Anticipating and defending against such adversarial strategies will be essen- tial to preserve the long-term relevance of biologically inspired detection. Pursuing these directions will allow Forensic Lens to mature into a more resilient, generalizable, and practically deployable solution—one that balances accuracy, efficiency, and interpretability in the continuing effort to safeguard digital trust.

8. REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[2] Amit Agarwal, Srikant Panda, Angeline Charles, Bhargava Kumar, Hitesh Patel, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, et al. Mvtamperbench: Evaluating robustness of vision-language models. *arXiv preprint arXiv:2412.19794*, 2024.

[3] Ahmad ALBarqawi, Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah, and NhatHai Phan. Vigtext: Deepfake image detection with vision- language model explanations and graph neural networks. *arXiv preprint arXiv:2507.18031*, 2025.

[4] Abdullah Alharbi, Wael Alosaimi, Mohd Nadeem, Hashem Alyami, Bader Alouffi, Ahmed Almulihi, Nafees Akhter Farooqui, Rafeeq Ahmed, and Raees Ahmad Khan. Novel 59-layer dense inception net-work for robust deepfake identification. *Scientific Reports*, 15(1):24159, 2025.

[5] Wasin Alkishri, Setyawan Widyarto, and Jabar H Yousif. Evaluating the effectiveness of a gan fingerprint removal approach in fooling deepfake face detection. *Journal of Internet Services and Information Security (JISIS)*, 14(1):85–103, 2024.

[6] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al. Deepfake media forensics: Status and future challenges. *Journal of Imaging*, 11(3):73, 2025.

[7] Joseph Bamidele Awotunde, Rasheed Gbenga Jimoh, Agbotiname Lucky Imoize, Akeem Tayo Abdulrazaq, Chun-Ta Li, and Cheng-Chi

Lee. An enhanced deep learning-based deepfake video detection and classifica- tion system. *Electronics*, 12(1):87, 2022.

[8] Saravana Balaji Balasubramanian, P Prabu, K Venkatachalam, Pavel Trojovsky`, et al. Deep fake detection using cascaded deep sparse auto-encoder for effective feature selection. *PeerJ Computer Science*, 8:e1040, 2022.

[9] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[10] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[11] Umur Aybars C¸ iftc¸i, Ilke Demir, and Lijun Yin. Deepfake source detection in a heart beat. *The Visual Computer*, 40(4):2733–2750, 2024.

[12] Sara Concas, Simone Maurizio La Cava, Giulia Orru`, Carlo Cuccu, Jie Gao, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli. Analysis of score-level fusion rules for deepfake detection. *Applied Sciences*, 12(15):7365, 2022.

[13] Hussain Dawood, Marriam Nawaz, Tahira Nazir, Ali Javed, Abdul Khader Jilani Saudagar, and Hatoon S AlSagri. Arnet: Integrating spatial and temporal deep learning for robust action recognition in videos. *Computer Modeling in Engineering & Sciences (CMES)*, 144(1), 2025.

[14] Shahad Eidan et al. Unmasking deepfakes: A systematic review of generation techniques and detection strategies. *Iraqi Journal of Intelligent Computing and Informatics (IJICI)*, 4(2):134–154, 2025.

[15] Yuan Gao, Xuelong Wang, Yu Zhang, Ping Zeng, and Yingjie Ma. Tem- poral feature prediction in audio–visual deepfake detection. *Electronics*, 13(17):3433, 2024.

[16] Muskan Garg. Towards mental health analysis in social media for low- resourced languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–22, 2024.

[17] Sergio Gonza´lez, Salvador Garc´ıa, Javier Del Ser, Lior Rokach, and Francisco Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64:205–237, 2020.

[18] Shivam Grover, Amin Jalali, and Ali Etemad. Segment, shuffle, and stitch: A simple layer for improving time-series representations. *Advances in Neural Information Processing Systems*, 37:4878–4905, 2024.

[19] David Gu¨era and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[20] Mahmudul Hasan, Sadia Ruhama, Sabrina Tajnim Sithi, Chowdhury Mohammad Mutamir Samit, and Oindrila Saha. Unmasking deep fakes: Leveraging deep learning for video authenticity detection. *arXiv preprint arXiv:2505.06528*, 2025.

[21] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakes detection based on heart rate estimation: Single-and multi-frame. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pages 255–273. Springer International Publishing Cham, 2022.

[22] N Jabbar and Ch Nadeem. Modeling & evaluating the performance of convolutional neural networks for classifying steel surface defects. *International Journal of Advanced*

Computer Science and Applications (IJACSA), 14(6):123–131, 2023.

[23] Muhammad Javed, Zhaohui Zhang, Fida Hussain Dahri, and Asif Ali Laghari. Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. *Electronics*, 13(15):2947, 2024.

[24] Wurood A Jbara, Noor Al-Huda K Hussein, and Jamila H Soud. Deepfake detection in video and audio clips: a comprehensive survey and analysis. *Mesopotamian Journal of CyberSecurity*, 4(3):233–250, 2024.

[25] Bachir Kaddar, Sid Ahmed Fezza, Wassim Hamidouche, Zahid Akhtar, and Abdenour Hadid. Hcit: Deepfake video detection using a hybrid model of cnn features and vision transformer. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[26] Sukhandeep Kaur, Mubashir Buhari, Naman Khandelwal, Priyansh Tyagi, and Kiran Sharma. Hindi audio-video-deepfake (hav-df): A hindi language-based audio-video deepfake dataset. *arXiv preprint arXiv:2411.15457*, 2024.

[27] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. Evalua- tion of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, pages 7–15, 2021.

[28] Janavi Khochare, Chaitali Joshi, Bakul Yenarkar, Shraddha Suratkar, and Faruk Kazi. A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*, 47(3):3447–3458, 2022.

[29] Aminollah Khormali and Jiann-Shiun Yuan. Add: Attention-based deepfake detection approach. *Big Data and Cognitive Computing*, 5(4):49, 2021.

[30] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020.

[31] Lukas Kroiß and Johannes Reschke. Deepfake detection of face images based on a convolutional neural network. *arXiv preprint arXiv:2503.11389*, 2025.

[32] Gihun Lee and Mihui Kim. Deepfake detection using the rate of change between frames based on computer vision. *Sensors*, 21(21):7367, 2021.

[33] Shengyin Li, Vibekananda Dutta, Xin He, and Takafumi Matsumaru. Deep learning based one-class detection system for fake faces generated by gan network. *Sensors*, 22(20):7767, 2022.

[34] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018.

[35] Chin-Yuan Lin, Jen-Chun Lee, Shuenn-Jyi Wang, Chung-Shi Chiang, and Chao-Lung Chou. Video detection method based on temporal and spatial foundations for accurate verification of authenticity. *Electronics*, 13(11):2132, 2024.

[36] Chi Liu, Tianqing Zhu, Yuan Zhao, Jun Zhang, and Wanlei Zhou. Dis- entangling different levels of gan fingerprints for task-specific forensics. *Computer Standards & Interfaces*, 89:103825, 2024.

[37] Xiaolong Liu, Yang Yu, Xiaolong Li, and Yao Zhao. Mcl: multimodal contrastive learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2803–2813, 2023.

[38] Priyanka Muruganandham, Govardhana Rajan Thangasamy, Sangeetha Jayaraman, and Rekha Dharmarajan. Lstm autoencoder based parallel architecture for deepfake audio detection with

dynamic residual encod- ing and feature fusion. *Scientific Reports*, 15(1):23514, 2025.

[39] Muhammad Yasir Nadeem Ch, Siddiqui and Sanaullah Manzoor. Media forensics and deepfake: A systematic survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(10):456– 465, 2023.

[40] Gourab Naskar, Sk Mohiuddin, Samir Malakar, Erik Cuevas, and Ram Sarkar. Deepfake detection using deep feature stacking and meta-learning. *Heliyon*, 10(4), 2024.

[41] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525, 2022.

[42] Abdul Qadir, Rabbia Mahum, Mohammed A El-Meligy, Adham E Ragab, Abdulmalik AlSalman, and Muhammad Awais. An efficient deepfake video detection using robust deep learning. *Heliyon*, 10(5), 2024.

[43] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4318–4327, 2020.

[44] Rimsha Rafique, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri. Deep fake detection and classification using error-level analysis and deep learning. *Scientific reports*, 13(1):7422, 2023. Hidayat Ur Rahman, Ch Nadeem, Sanaullah Manzoor, F Najeeb, Muhammad Yasir Siddique, and RA Khan. A comparative analysis of machine learning approaches for plant disease identification. *Advance- ments in Life Sciences*, 4(4):120–126, 2017.

[45] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[46] Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaoheng Ni, Anurag Kumar, Marvin Sach, Yihui Fu, Wei Wang, et al. Interspeech 2025 urgent speech enhancement challenge. *arXiv preprint arXiv:2505.23212*, 2025.

[47] Ousama A Shaaban and Remzi Yildirim. Audio deepfake detection using deep learning. *Engineering Reports*, 7(3):e70087, 2025.

[48] Misaj Sharafudeen and Vinod Chandra SS. Frequency forensics for deep fake face detection using dual residual networks. *Multimedia Tools and Applications*, pages 1–26, 2025.

[49] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, and Peyman Najafirad. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, 4:100217, 2022.

[50] Stuart A Thompson. How ‘deepfake elon musk’became the internet’s biggest scammer. *New York Times*, 14, 2024.

[51] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinfor- mation: Exploring the impact of synthetic political video on de-ception, uncertainty, and trust in news. *Social media+ society*, 6(1):2056305120903408, 2020.

[52] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020.

[53] Yuxi Wang, Yikang Wang, Qishan Zhang, Hiromitsu Nishizaki, and Ming Li. Vcapav: A video-caption based audio-visual deepfake detection dataset. In *Proc. Interspeech 2025*, pages 3908–3912, 2025.

[54] Kevin Warren, Daniel Olszewski, Seth Layton, Kevin Butler, Carrie Gates, and Patrick Traynor. Pitch imperfect: Detecting audio deepfakes through acoustic prosodic analysis. *arXiv preprint arXiv:2502.14726*, 2025.

[55] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion prob- abilistic modeling for video generation. *Entropy*, 25(10):1469, 2023.

[56] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14800–14809, 2021.

[57] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2929– 2939, 2021.