International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

Deep Guard: A Multimodal AI-Based Deepfake Detection System for Manipulated Images, Videos, and Audio

V. S. Dhande[1], M. R. Shaikh[2], Abhiraj Abhale[3], Sidharth Ambhore[4], Om Chakranarayan[5], Bhushan Bendke[6]

[2,3,4,5,6]*Department of Computer Science and Engineering Students, Group PG-07*
[1]*Department of Computer Science and Engineering, Project Guide*

***Abstract*—The rapid proliferation of deep learning-based media synthesis techniques, commonly referred to as deepfakes, poses a significant threat to digital trust, public discourse, and indi- vidual privacy. Existing detection systems often address only a single modality—either visual or auditory—leaving them vulner- able when adversaries combine manipulated video frames with cloned speech. In this paper, we propose DeepGuard, a unified multimodal deepfake detection framework that jointly analyzes images, video sequences, and audio streams. DeepGuard in- tegrates a Convolutional Neural Network (CNN) branch for spatial artifact extraction, a Vision Transformer (ViT) branch for long-range temporal dependencies, and a ResNet-based spec- trogram classifier for audio anomaly detection. A late-fusion module aggregates modality-specific confidence scores to pro- duce a final authenticity verdict.**

***Keywords*—Deepfake Detection, Convolutional Neural Networks, Vision Transformer, Multimodal Learning, Audio Spoofing, Face- Forensics++, Media Forensics, Face Manipulation, ASVspoof, Deep Learning.**

I. Introduction

A. What Are Deepfakes?

Deepfakes are hyper-realistic synthetic media generated by deep neural networks—most notably Generative Adversarial Networks (GANs) [1] and Variational Autoencoders—that seamlessly replace or alter a person's likeness, voice, or both. The term was coined on the Reddit platform in 2017, where users began publishing face-swap videos created with early encoder–decoder architectures [2]. Since then, the quality of generated content has improved dramatically; modern tools such as DeepFaceLab, FaceSwap, and commercial services can synthesize near-photorealistic videos in minutes on consumer-grade hardware.

B. Problems Caused by Deepfake Technology

The misuse of deepfake technology carries far-reaching societal consequences:

- *Disinformation and Political Manipulation:* Fabricated videos of political figures making inflammatory statements can destabilize elections and erode public trust [3].
- *Non-Consensual Intimate Imagery:* The majority of deepfake content online involves non-consensual porno- graphic material, causing severe psychological harm to victims [4].
- *Financial Fraud:* Voice-cloning attacks have been used to impersonate executives and authorize fraudulent wire transfers, resulting in multi-million-dollar losses [5].
- *Evidence Tampering:* The prospect of fabricated audiovisual evidence threatens the integrity of legal pro- ceedings and investigative journalism [6].

C. Importance of Detecting Manipulated Media

Timely and accurate detection of deepfakes is essential to preserve the authenticity of digital evidence, protect individual reputations, and safeguard democratic institutions. Automated detection tools must operate at scale across heterogeneous platforms—social networks, messaging applications, broadcast media—where manual forensic review is infeasible. Moreover, as generative models continue to improve, detectors must be adaptive, generalizing to unseen manipulation methods rather than overfitting to artifacts of a particular generator [7].

This paper makes the following contributions:

1. A unified three-branch multimodal architecture (CNN + ViT + audio classifier) for simultaneous image, video, and speech forgery detection.
2. A late-fusion confidence aggregation strategy that as- signs modality-specific reliability weights.
3. Comprehensive evaluation on four publicly available benchmarks, surpassing current state-of-the-art

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

II. Literature Review

A. Early Forensic Approaches

Prior to the deep learning era, media forensics relied on handcrafted features such as Benford's Law statistics, double JPEG compression artifacts, and chromatic aberration analysis [8]. While effective for traditional photo editing, these methods proved inadequate against GANs, which synthesize images from noise without retaining classical compression footprints.

B. CNN-Based Detection

Meso-4 and MesoInception-4 [9] were among the first dedicated CNN architectures for deepfake detection, operating on raw pixel patches to identify mesoscopic properties of generated faces. Rossler et al. [7] proposed using XceptionNet—an extreme version of Inception with depthwise separable convolutions—as a backbone and introduced the FaceForensics++ benchmark. Li et al. [10] exploited face-warping artifacts inherent in many face-swap pipelines, achieving strong performance on both FaceForensics++ and Celeb-DF. Stehouwer et al. proposed attention-based CNNs that guide the network towards facial boundaries and ocular regions where blending artifacts frequently appear [11].

C. Transformer-Based Detection

Inspired by the success of the Vision Transformer (ViT) [12] for image classification, several works have applied attention mechanisms to deepfake detection. Zhao et al. [13] introduced a multi-attentional approach that captures fine-grained regional inconsistencies across texture, color, and geometry. FTCN [14] uses fully temporal convolution networks to model cross-frame consistency. More recently, CLIP-based [15] vision-language models have been adapted as zero-shot deepfake detectors, leveraging large-scale pretraining to generalize across unseen generators.

D. Audio Deepfake Detection

The ASVspoof challenge series [16, 17] has been instrumental in advancing audio anti-spoofing research. LightCNN (LCNN) [18] demonstrated that CNNs applied to log-magnitude spectrograms are competitive with GMM-UBM baselines. RawNet2 [19] operates directly on raw waveforms using sinc filters, obviating manual feature engineering. Tak et al. [20] proposed RawBoost data augmentation to improve robustness against channel and codec variability. Graph Attention Networks have also been applied to model phoneme-level spoofing cues [21].

E. Multimodal Approaches

Despite the prevalence of audio-visual deepfakes, comparatively few works jointly model both modalities. Zhou et al. [22] proposed a joint audio-visual deepfake detection framework exploiting cross-modal synchrony, observing that real videos exhibit tighter lip-sync coherence than forgeries. Our work extends this line by adding a dedicated ViT branch for long-range temporal reasoning and a learnable late-fusion weighting mechanism, enabling more reliable predictions across diverse forgery types.

III. Proposed Methodology

A. System Overview

DeepGuard accepts three types of input: static images, video clips, and raw audio waveforms. The pipeline consists of four stages: (1) data preprocessing and face/speech segmentation, (2) modality-specific feature extraction, (3) per-branch classification, and (4) multimodal late fusion.

B. Data Preprocessing

1) Frame Extraction and Face Alignment: For video inputs, frames are extracted at 25 fps using FFmpeg. RetinaFace [23] is employed for face detection, producing a bounding box and five facial landmarks per detection. Detected face crops are aligned to a canonical 224×224 pixel representation via an affine transformation based on the eye and mouth landmark coordinates. Crops are normalized to zero mean and unit variance using ImageNet statistics for compatibility with pre-trained backbone weights.

2) Spectrogram Generation: Raw audio waveforms are resampled to 16 kHz and segmented into 4-second windows with 50% overlap. Each window is converted to a log-scale mel-spectrogram with 128 mel filterbanks, a window size of 25 ms, and a hop length of 10 ms. The resulting 128×400 2-D spectrogram image is fed to the audio classification branch.

3) Data Augmentation: To improve generalization, training-time augmentation is applied independently per modality. Visual augmentations include random horizontal flipping, rotation ($\pm 15^{\circ}$), color jitter (brightness and contrast ± 0.2), and Gaussian blur. Audio augmentations include additive noise (SNR 10–30 dB), room impulse response convolution, codec simulation (MP3 at 32–128 kbps), and the RawBoost data augmentation strategy [20].

C. Model Architecture

1) CNN Visual Branch: The CNN branch uses EfficientNet-B4 [24] pre-trained on ImageNet as a feature extractor. The final global average pooling layer produces a 1792-dimensional embedding, which is passed through two fully connected layers (FC-512 → FC-256, both with BatchNorm and ReLU) and a sigmoid output node for binary classification.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

2) Vision Transformer Branch: The ViT branch uses a ViT-B/16 [12] model pre-trained on ImageNet-21k. Each 224×224 face crop is divided into 196 non-overlapping 16×16 patches. Patch embeddings are passed through 12 transformer encoder blocks. The classification token (`[CLS]`) representation after the final block is projected via a linear head to a binary logit. This branch captures global and long-range spatial relationships that local CNNs may miss.

3) Audio Classification Branch: The audio branch uses a ResNet-34 backbone applied to the mel-spectrogram images. We replace the standard first convolutional layer with a 3×3 kernel (stride 1) to better preserve high-frequency spectral details. The 512-dimensional penultimate feature vector is passed through a FC-128 layer and a binary output head.

D. Late Fusion Module

The three branches produce scalar confidence scores s_{CNN}, s_{ViT}, and s_{audio} in $[0, 1]$. A learnable weighted average aggregates these scores:

$$s_{\text{fused}} = w_1 s_{\text{CNN}} + w_2 s_{\text{ViT}} + w_3 s_{\text{audio}}, \quad w_i \geq 0, \quad \sum_{i=1}^{3} w_i = 1. \tag{1}$$

The weights $\{w_i\}$ are optimized end-to-end using the joint training loss. When only a single modality is available, the corresponding branch is activated and fusion weights are renormalized. The final prediction is:

$$\hat{y} = \mathbf{1}[s_{\text{fused}} \geq \tau], \tag{2}$$

where $\tau = 0.5$ by default (tunable on a held-out validation set).

E. Training Process

All branches are trained jointly using binary cross-entropy loss. The AdamW optimizer [27] is used with an initial learning rate of 2×10^{-4}, a cosine annealing schedule with warm restarts, and a weight decay of 10^{-4}. Training is conducted for 40 epochs on 4 NVIDIA A100 GPUs with a per-GPU batch size of 32, using mixed-precision (FP16) training to reduce memory overhead. Early stopping with a patience of 8 epochs is applied based on validation AUC.

IV. System Architecture

The DeepGuard system comprises five interconnected modules:

1. *Input Ingestion Module:* Accepts raw media files (JPEG/PNG images, MP4/AVI videos, WAV/FLAC au- dio). A media type detector routes inputs to the appro- priate preprocessing sub-pipeline.
2. *Preprocessing Module:* Performs face detection and alignment for visual inputs (using RetinaFace) and mel- spectrogram generation for audio inputs, as described in Section III.
3. *Feature Extraction Module:* Three parallel sub-networks—EfficientNet-B4, ViT-B/16, and ResNet-34— independently encode modality-specific representations, each initialized with ImageNet pre-trained weights and fine-tuned on deepfake benchmark data.
4. *Classification & Fusion Module:* Per-branch binary classifiers produce confidence scores aggregated by the learnable late-fusion module (Eq. 1) to yield a unified forgery probability.
5. *Result Generation Module:* Outputs a human-readable verdict (REAL/FAKE), a confidence percentage, a saliency map (Grad-CAM [25]) highlighting regions influencing the decision, and a JSON metadata record for downstream integration.

V. Dataset Description

A. FaceForensics++ (FF++)

FaceForensics++ [7] is a large-scale video forgery dataset containing 1,000 original YouTube videos manipulated by four methods: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Videos are available at three quality levels: raw, high-quality (HQ, CRF 23), and low- quality (LQ, CRF 40). We train and evaluate on the HQ split following standard protocol, using the official 720/140/140 train/validation/test split.

TABLE 1:
DeepGuard Module Summary

Module	Key Components
Input Ingestion	Media type detector, format parser
Preprocessing	RetinaFace, affine alignment, FFmpeg, mel-spectrogram
Feature Extraction	EfficientNet-B4, ViT-B/16, ResNet-34
Fusion	Learnable weighted average, sigmoid
Result Generation	Grad-CAM, JSON output, verdict label

B. Celeb-DF v2

Celeb-DF v2 [10] addresses the quality gap of earlier datasets by providing 5,639 high-quality celebrity deepfake videos synthesized with an improved face-swap pipeline that reduces visible blending artifacts. The dataset includes 590 real videos and 5,049 forgeries spanning 59 celebrities. Celeb-DF v2 is notably challenging because its forgeries are of commercial-grade quality.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

C. Deepfake Detection Challenge (DFDC)

The DFDC dataset [26], released as part of the Facebook AI challenge, comprises over 128,000 video clips of 3,426 consenting participants. Forgeries were produced using eight generation methods including face swap, puppet-master, and expression transfer techniques. The dataset is highly diverse in terms of lighting, pose, and background, making it one of the most challenging benchmarks.

D. ASVspoof 2019

ASVspoof 2019 [17] is the standard benchmark for audio anti-spoofing, containing Logical Access (LA) and Physical Access (PA) conditions. The LA partition includes 19 text-to-speech and voice conversion systems, while PA simulates replay attacks in various acoustic environments. Performance is measured using the Equal Error Rate (EER) metric. We use the LA condition for voice-clone detection, consistent with prior work.

VI. Experimental Results

A. Evaluation Metrics

We report Accuracy (ACC), Precision (P), Recall (R), F1-Score (F1), and Area Under the ROC Curve (AUC). All metrics are computed at the video/clip level by averaging frame-level predictions.

B. Performance on Individual Datasets

C. Comparison with State-of-the-Art

D. Confusion Matrix Analysis

On the combined test set (all four datasets), the model produced:

- *True Positives (TP):* 9,421 (correctly identified fakes)
- *True Negatives (TN):* 3,104 (correctly identified reals)

TABLE 2:
Per-Dataset Detection Performance (%)

Dataset	ACC	P	R	F1
FF++ (HQ)	99.1	99.0	99.2	99.1
Celeb-DF v2	96.8	96.5	96.3	96.4
DFDC	95.7	95.1	94.8	95.0
ASVspoof (LA)	97.9	97.8	96.9	97.3
Average	**97.4**	**97.1**	**96.8**	**96.9**

TABLE 3:
Comparison with Baseline Methods (FF++ HQ, ACC %)

Method	Venue	ACC (%)
MesoNet [9]	WIFS'18	83.1
XceptionNet [7]	ICCV'19	95.7
Multi-Attention [13]	CVPR'21	97.6
FTCN [14]	ICCV'21	98.8
DeepGuard (Ours)	—	**99.1**

- *False Positives (FP):* 276 (real media flagged as fake)
- *False Negatives (FN):* 322 (fakes missed by the model)

The low false-negative rate is critical for security-critical applications where missed deepfakes carry higher risk than false alarms.

E. Ablation Study

TABLE 4:
Ablation Study: Contribution of Each Branch (ACC %)

Configuration	FF++	Celeb-DF	DFDC
CNN only	95.7	91.2	88.4
ViT only	96.9	93.5	90.1
CNN + ViT	98.5	95.8	93.7
CNN + Audio	97.1	93.0	91.5
Full Fusion	**99.1**	**96.8**	**95.7**

The ablation confirms that each branch contributes meaningfully, and that late-fusion consistently outperforms any single-modality configuration.

VII. Discussion

A. Strengths

- *Cross-Modal Robustness:* By fusing visual and acous- tic signals, DeepGuard is resilient to attacks that replace only one modality, a common adversarial strategy.
- *Generalization:* Pre-training on large-scale datasets (ImageNet-21k for ViT, ASVspoof for audio) provides a strong initialization that improves cross-dataset transfer.
- *Interpretability:* Grad-CAM visualizations allow ana- lysts to understand which facial regions or spectral bands triggered a detection, supporting human oversight.
- *Modality Flexibility:* The system degrades gracefully when one modality is absent (e.g., silent video), activat- ing only the relevant branches.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 03, March 2026)

B. Limitations

- *Computational Cost:* Running three deep networks in parallel is computationally expensive. Inference on a single V100 GPU takes approximately 480 ms per video second, limiting real-time applicability.
- *Adversarial Robustness:* Like most learning-based de- tectors, DeepGuard is susceptible to adversarial pertur- bations specifically crafted to fool the model.
- *Compression Sensitivity:* Heavily compressed videos (CRF ≥ 40) reduce detection accuracy by up to 7% compared to raw-quality inputs.
- *Dataset Bias:* Training predominantly on Western celebrity datasets may reduce performance on media featuring other demographics and cultural contexts.

VIII. Conclusion

This paper presented DeepGuard, a multimodal AI-based deepfake detection system that simultaneously analyzes image, video, and audio streams. By combining EfficientNet-B4, ViT-B/16, and ResNet-34 branches with a learnable late-fusion module, DeepGuard achieves state-of-the-art detection accuracy across four benchmark datasets, reaching an average accuracy of 97.4% and an F1-score of 96.9%. Ablation studies confirm that multimodal fusion consistently outperforms single-modality baselines, underscoring the complementary nature of visual and acoustic forgery cues. While limitations remain—particularly regarding real-time inference speed and adversarial robustness—the results demonstrate that unified multimodal architectures represent a promising direction for next-generation media forensics. As deepfake generation technology continues to advance, the research community must prioritize adaptive, generalizable detection systems that can keep pace with emerging synthesis methods.

IX. Future Work

Real-Time Detection: We plan to explore knowledge distillation [28] and neural architecture search to produce lightweight student models capable of real-time inference on edge devices and mobile platforms, targeting sub-50 ms latency per video second.

Multimodal Detection Improvements: Future work will investigate early-fusion and cross-attention fusion mechanisms that explicitly model temporal lip-speech synchrony, facial micro-expression dynamics, and audio-visual emotion consistency, potentially providing richer forgery signals than independent late fusion.

Adversarial Robustness: We intend to incorporate adversarial training with PGD [29] and certification-based defenses to harden the system against white-box and black-box adversarial perturbations.

Larger and More Diverse Datasets: Collecting and annotating large-scale datasets spanning multiple languages, ethnicities, and recording conditions is critical. Synthetic data augmentation using controllable diffusion models may also help bridge the coverage gap.

Explainability and Forensic Reports: Integrating ad- vanced explainability methods—such as SHAP [30] values and concept activation vectors—will enable the generation of standardized forensic reports suitable for legal proceedings.

Cross-Platform Deployment: We aim to deploy DeepGuard as a browser extension and API service to support fact-checkers and journalism organizations with automated pre-publication media verification.

References

[1] I. Goodfellow et al., "Generative adversarial nets," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2014, pp. 2672–2680.

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131–148, 2020.

[3] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," Calif. Law Rev., vol. 107, pp. 1753–1820, 2019.

[4] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The state of deepfakes: Landscape, threats, and impact," Deeptrace Labs, Tech. Rep., 2019.

[5] C. Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," Wall Street J., Aug. 2019.

[6] B. Paris and J. Donovan, "Deepfakes and cheap fakes," Data & Society Research Institute, Tech. Rep., 2019.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, 2019, pp. 1–11.

[8] H. Farid, "Image forgery detection," IEEE Signal Process. Mag., vol. 26, no. 2, pp. 16–25, 2009.

[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE WIFS, 2018.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in Proc. IEEE/CVF CVPR, 2020, pp. 3207–3216.

[11] J. Stehouwer, H. Dang, J. Liu, F. Liu, and X. Wan, "Detection of fake and fraudulent faces via neural memory networks," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1467–1478, 2021.

[12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[13] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, 2021, pp. 2185–2194.

[14] J. Zheng et al., "Exploring temporal coherence for more general video face forgery detection," in Proc. IEEE/CVF ICCV, 2021, pp. 15044–15054.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.

[16] Z. Wu et al., "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in Proc. Interspeech, 2015, pp. 2037–2041.

[17] A. Nautsch et al., "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," IEEE Trans. Biometrics Behav. Identity Sci., vol. 3, no. 2, pp. 252–265, 2021.

[18] G. Lavrentyeva et al., "Audio replay attack detection with deep learning frameworks," in Proc. Interspeech, 2017, pp. 82–86.

[19] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in Proc. IEEE ICASSP, 2021, pp. 6369–6373.

[20] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Raw-Boost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in Proc. IEEE ICASSP, 2022, pp. 6382–6386.

[21] J. Jung et al., "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in Proc. IEEE ICASSP, 2022, pp. 6367–6371.

[22] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Joint audio-visual deepfake detection," in Proc. IEEE/CVF ICCV, 2021, pp. 14800–14809.

[23] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in Proc. IEEE/CVF CVPR, 2020, pp. 5203–5212.

[24] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE/CVF ICCV, 2017, pp. 618–626.

[26] B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.