OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21610315

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Articles

Volume-06|Issue-04|2026

Multi-Modal Deepfake Detection System Using Hybrid Deep Learning on Visual and Audio Features

Nandana K Gowda[1], Hemavathi P[2]

[1,2]Department of Computer Science & Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Gowda, N. & Hemavathi, P. (2026) Multi-Modal Deepfake Detection System Using Hybrid Deep Learning on Visual and Audio Features. *Indiana Journal of Multidisciplinary Research, 6*(4), 393-399.

Abstract: Deepfake technology, driven by generative models such as GANs and diffusion architectures, has enabled the creation of highly realistic manipulated media capable of deceiving both visual and auditory perception. Such forgeries pose significant risks to identity verification, digital forensics, and public trust. Existing detection methods are often restricted to single modalities or handcrafted features, limiting their robustness against high-quality cross-modal deepfakes. This work presents a multi-modal deepfake detection framework that integrates visual and audio analysis. The system employs a ResNet-based backbone for spatial feature extraction, Bi-LSTM encoders for temporal modeling in videos, and a 2D CNN with Random Forest classification for speech-based anomaly detection. An attention-based late-fusion strategy combines outputs to generate reliable predictions with calibrated confidence scores. Experimental results demonstrate high performance, with training accuracy reaching 98% and validation accuracy of 92% on benchmark datasets. The system is deployed as a Flask-based web application, enabling real-time detection and interpretable outputs.

Keywords: Deepfake Detection, Multi-Modal Learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), ResNet

INTRODUCTION

The term **deepfake** refers to synthetic media produced by advanced deep learning models—most notably Generative Adversarial Networks (GANs) and diffusion-based architectures—that can manipulate or fabricate video and audio to closely resemble authentic recordings. While deepfakes have legitimate applications in domains such as film production and virtual reality, their misuse presents serious risks to individuals and society. Malicious deepfakes can spread misinformation, damage reputations, facilitate identity fraud, and undermine public trust in digital media authenticity.

Recent improvements in generative modelling have diminished the effectiveness of many traditional detection approaches. Early methods relied on handcrafted features and classical machine learning classifiers (e.g., support vector machines) to identify manipulation traces, but these methods struggle with high-resolution and temporally consistent deepfakes produced by modern generators and therefore suffer from high false-positive and false-negative rates.

Many detectors have also been unimodal (visual only or audio only), which limits their ability to identify cross-modal inconsistencies: a visual-only system may miss mismatched lip movements, whereas an audio-only system may overlook subtle visual anomalies in skin texture or eye blinking.

In addition, several existing tools are platform-restricted or computationally expensive, reducing their suitability for real-time or widely accessible deployments.

To address these limitations, this work proposes a hybrid multi-modal detection framework that unifies visual and audio analysis. Visual features are extracted using deep convolutional backbones (e.g., ResNet) with temporal dependencies modelled by bi-directional LSTM encoders. Audio is analyzed using spectral feature extraction (e.g., MFCCs) and a 2D convolutional pipeline followed by an ensemble classifier. Multimodal representations are combined via an attention-based late-fusion strategy to exploit cross-modal cues and improve robustness to high-quality forgeries.

The proposed system is deployed as a Flask-based web application that accepts media uploads and returns detection outcomes with calibrated confidence scores. Empirical results indicate strong performance improvements over unimodal baselines, demonstrating the value of joint audio–visual reasoning for deepfake detection.

RELATED WORKS

Early deepfake detection research focused on handcrafted features and shallow classifiers such as SVMs, seeking visual inconsistencies (e.g., unnatural blinking or facial warping). These approaches were limited in generalization and robustness when confronted with GAN-generated high-quality forgeries [1].

***Corresponding Author:** Nandana K Gowda

With the rise of deep learning, convolutional neural network (CNN)-based methods became predominant. Architectures such as ResNet and DenseNet have been widely adopted to learn discriminative representations from facial images and video frames. Hybrid CNNs with constrained convolutions have been effective at suppressing editing traces produced by manipulation tools like DeepFake and Face2Face [1,2]. Gabor filter-based CNNs and attention-augmented networks further improved sensitivity to subtle texture and expression artifacts while maintaining computational efficiency [4]. Nonetheless, purely visual CNN methods still face challenges across diverse datasets and compression levels.

To improve generalization, attention-driven and residual frameworks were investigated. Methods combining Mask R-CNN with residual attention have localized manipulation artifacts (for example, discrepancies in eye reflections) [5]. Meta-learning solutions using block-shuffling and pair-attention strategies were proposed to enhance cross-dataset adaptability [3].

Frequency- and noise-based detectors targeted residual artifacts invisible in pixel space. Approaches leveraging SRNet, Laplacian blur variance, and local frequency-guided networks detected manipulations through distortions in noise and spectral patterns, increasing robustness to subtle forgeries [2,9]. These works highlighted the importance of fusing spatial and frequency-domain cues.

More recent architectures include graph-based and transformer models. Graph convolutional networks model facial landmarks as nodes to analyze relational facial dynamics, while Vision Transformers with relevancy maps identify suspicious regions within frames [12]. Although these approaches often deliver improved resilience to compression and blur, they typically incur higher computational costs.

Multimodal systems that fuse visual and audio features have also emerged, capturing cross-modal inconsistencies such as lip-sync mismatches and unnatural prosody, and showing improved detection accuracy over unimodal baselines [7].

A complementary line of progress has been dataset development. Large-scale, diverse datasets (e.g., eKYC-DF) provide extensive training material and better evaluation conditions for identity-focused scenarios [18]. Other datasets and frameworks (e.g., mask-robust datasets and federated learning schemes) address privacy-preserving detection and domain shifts caused by real-world conditions [8,14].

Ensemble and hybrid approaches combining multiple backbones (ResNet, DenseNet, co-occurrence matrices) and voting mechanisms have shown gains in precision and adaptability to novel manipulations [16]. Hybrid GAN–ResNet systems with attention modules leverage both synthetic augmentation and feature extraction to improve robustness [17].

Altogether, prior work underscores that while CNNs and attention mechanisms have advanced deepfake detection, persistent challenges remain in cross-dataset generalization, real-time applicability, and resistance to novel generation techniques. Emerging directions—multimodal fusion, graph-based reasoning, transformer models, and federated learning—point toward more comprehensive and resilient detection systems.

PROPOSED METHODOLOGY

The proposed methodology introduces a comprehensive multi-modal deepfake detection framework that integrates visual and audio analysis to improve robustness against sophisticated manipulations. The architecture is organized into cooperating modules that process inputs in a complementary manner to ensure reliable detection.

A visual backbone based on ResNet residual blocks extracts spatial descriptors from full frames and focused local regions (e.g., mouth and eyes). Temporal dependencies are modelled by a temporal encoder composed of stacked bi-directional LSTM layers and temporal convolutional blocks to expose inconsistencies in lip-sync, blinking, and facial dynamics.

For audio, time–frequency representations (MFCCs, log-mel spectrograms, pitch contours, spectral contrast) are computed using Librosa and fed to a 2D CNN-based audio encoder that produces compact embeddings capturing timbre, phase artifacts, and synthesis traces. In parallel, a lightweight Random Forest classifier operates on engineered audio statistics as an orthogonal decision signal.

Outputs from the visual and audio pipelines are fused by an attention-based late-fusion module, which aggregates modality embeddings and produces calibrated confidence scores via fully connected layers and softmax with temperature scaling.

Dataset Description

The system is trained and evaluated on large-scale deepfake benchmarks. For example, the Deepfake Detection Challenge (DFDC) dataset contains extensive manipulated and genuine videos. Note that DFDC training data were distributed in large archives (totaling hundreds of gigabytes), and access required participant registration and compliance with competition rules. Training commonly uses FaceForensics++, Celeb-DF, and DFDC for breadth across manipulation types.

System Architecture

***Corresponding Author:** Nandana K Gowda

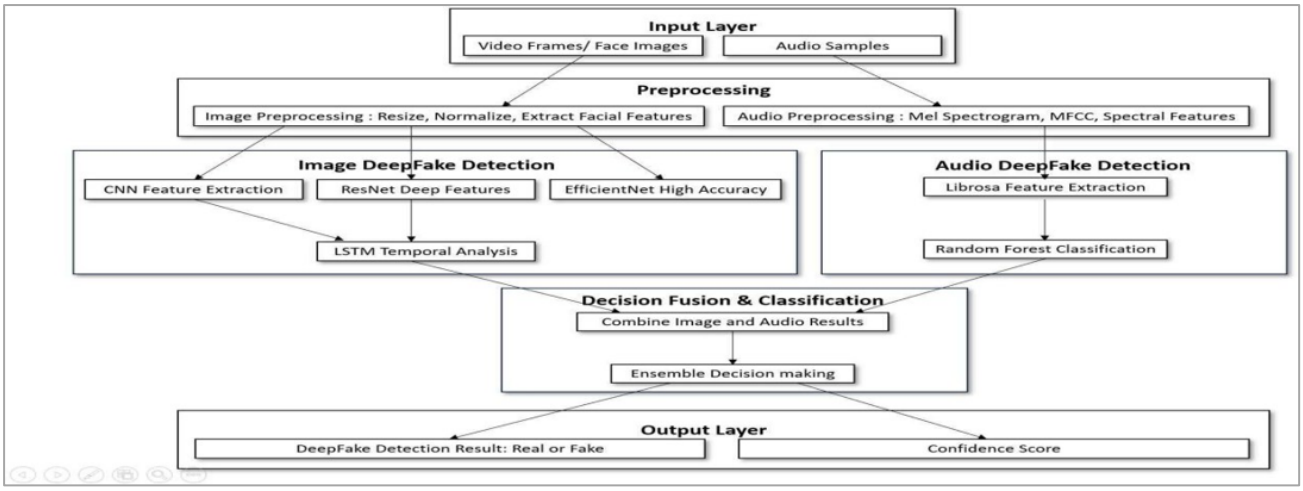

Fig. 1. Proposed multi-modal deepfake detection system architecture

Figure 1 depicts the five-layer architecture: input, preprocessing, detection modules, decision fusion, and output.

Input layer: Accepts video frames/face images for visual analysis and raw audio samples for speech analysis.

Preprocessing: Visual preprocessing performs resizing, normalization, and face-region extraction. Audio preprocessing computes Mel spectrograms, MFCCs, and other spectral descriptors.

Image-based detection: A CNN backbone (e.g., ResNet-50 or EfficientNet) extracts spatial features; stacked LSTM layers capture temporal irregularities such as inconsistent blinking or lip-sync errors.

Audio-based detection: Spectral/audio features extracted via Librosa are processed by a 2D CNN encoder and an auxiliary Random Forest classifier trained on engineered acoustic statistics.

Decision fusion: Modality outputs and confidences are combined using an attention-based late-fusion mechanism and ensemble rules to yield the final prediction.

Output layer: Returns a binary label (real/fake) together with a calibrated confidence score for interpretability.

IMPLEMENTATION DETAILS

The system is implemented as a modular pipeline using standard deep learning and signal-processing libraries.

Input Layer

Inputs supported: images, videos, and audio.

Videos are decoded and sampled into frames using OpenCV. Audio is resampled to 16 kHz for consistent feature extraction. Frames are typically extracted at a fixed frame rate (e.g., 25 fps) to balance temporal coverage and computation.

Preprocessing

Visual preprocessing: Frames are resized to **224 × 224** and normalized to pre-trained backbone statistics. Face regions are localized with MTCNN (or an equivalent detector) to focus analysis on relevant regions.

Audio preprocessing: Librosa is used to compute Mel spectrograms, MFCCs, chroma features, spectral contrast, and pitch contours. These are used as inputs to the audio encoder and the Random Forest feature set.

Image-Based Deepfake Detection

The image pipeline combines spatial and temporal extractors.

- **CNN feature extraction:** ResNet-50 (pretrained on ImageNet) is fine-tuned on deepfake datasets (e.g., FaceForensics++). EfficientNet is available as a lightweight alternative.
- **Temporal analysis:** Frame embeddings from the CNN are passed to stacked LSTM layers to capture sequential dynamics and reveal temporal inconsistencies.

Audio-Based Deepfake Detection

The audio module converts raw audio into spectral inputs (MFCCs, Mel spectrograms). These are processed by a 2D CNN encoder, producing embeddings. Separately, a Random Forest classifier (e.g., 500 trees) is trained on engineered audio statistics to provide a complementary decision signal robust to acoustic variations.

Decision Fusion and Classification

Visual and audio subsystems produce independent predictions and confidence values. The fusion module supports multiple ensemble strategies:

- Majority voting across subsystems.
- Weighted averaging of confidence scores with modality reliability weights.
- Threshold-based rules to reduce false positives.

***Corresponding Author:** Nandana K Gowda

An attention-based late-fusion block is applied to modality embeddings before the final classifier to exploit cross-modal interactions.

Output Layer and Deployment
Final outputs:
- Binary decision (real/fake).
- Confidence score in **[0,1]**.

The system is deployed as a Flask web application:
- Web UI for media upload and result presentation.
- Flask backend orchestrates preprocessing and routes data to detection modules.
- Results and metadata are logged in a backend database for audit and forensic review.

Training and Optimization
Training uses FaceForensics++, Celeb-DF, DFDC, and related datasets with augmentation (compression, additive noise, resolution scaling) to improve generalization.

Optimization details:
- CNN and LSTM modules: Adam optimizer with categorical cross-entropy, learning-rate scheduling, and early stopping.
- Random Forest: 500 trees tuned for a balance of accuracy and inference speed.

System Integration
The pipeline is containerized with Docker for reproducibility and deployed on GPU-enabled servers. CNN and LSTM workloads use GPU acceleration through TensorFlow or PyTorch, while the Random Forest runs efficiently on the CPU. The modular design enables independent updates to each component as improved techniques become available.

RESULTS

The proposed multi-modal system was evaluated on image, video, and audio samples to validate cross-modal detection capability. The results below report model outputs (label and confidence score) produced by the integrated visual and audio pipelines.

Fig. 2. Detection of a real image. The face region was localized with MTCNN, processed by a ResNet-50 backbone, and classified as **Real** with a high confidence score.

Figure 2 shows a genuine image correctly classified as **Real** after face localization and ResNet-50 feature extraction. The bounding box indicates the region used for classification.

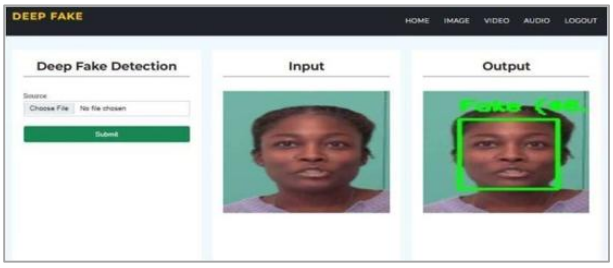

Fig. 3. Detection of a fake image. Spatial artifacts and temporal transition inconsistencies resulted in a Fake label.

Figure 3 presents a manipulated face image detected as **Fake** by the CNN–LSTM pipeline, demonstrating sensitivity to texture artifacts and sequence-level anomalies.

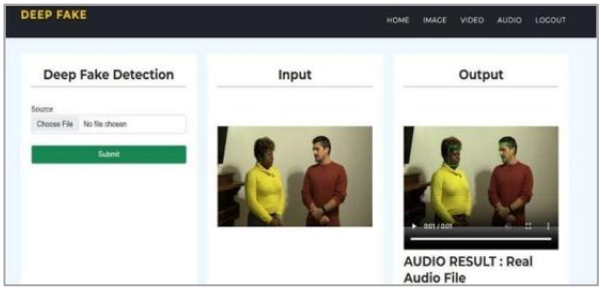

Fig. 4. Video deepfake detection. Frame-level predictions are aggregated to produce a single video-level decision.

As illustrated in Figure 4, videos are decomposed into frames, processed with ResNet-50 and Bi-LSTM, and aggregated to a video-level label, enabling detection of lip-sync and blinking inconsistencies.

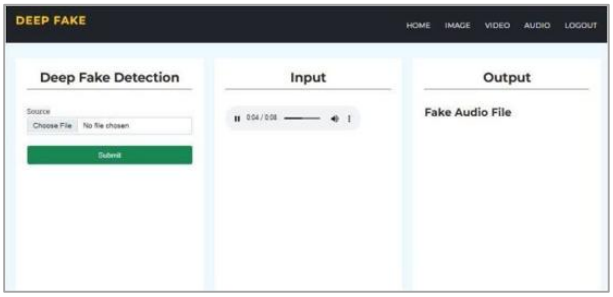

Fig. 5. Detection of a fake audio sample using spectral features and a Random Forest classifier (500 trees).

Figure 5 shows a synthesized voice sample flagged as Fake by the audio pipeline (MFCCs, spectral features → Random Forest), highlighting detectable frequency and harmonic irregularities.

***Corresponding Author:** Nandana K Gowda

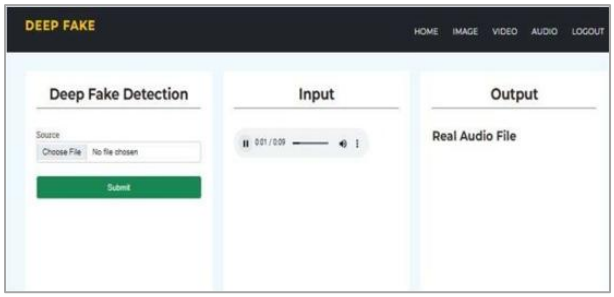

Fig. 6. Detection of a real audio sample correctly classified as Real.

Figure 6 demonstrates correct classification of an authentic audio recording, confirming the audio module's ability to distinguish natural harmonic and prosodic patterns.

Training and Validation Performance

Model learning behaviour was analysed by tracking loss and accuracy across epochs. Representative training curves are shown in Figures 7–10.

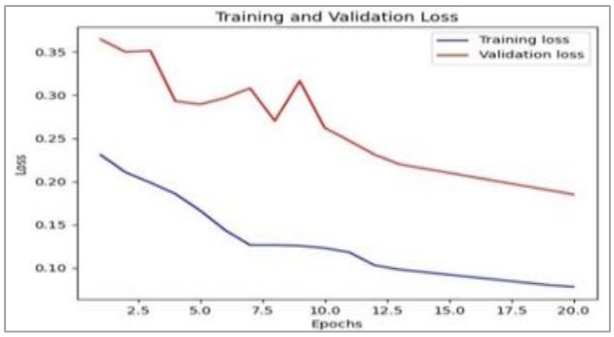

Fig. 7. Training and validation loss over 20 epochs.

Figure 7 shows a steady decline in training loss, with validation loss decreasing overall but exhibiting minor fluctuations, indicating effective learning with some validation instability.

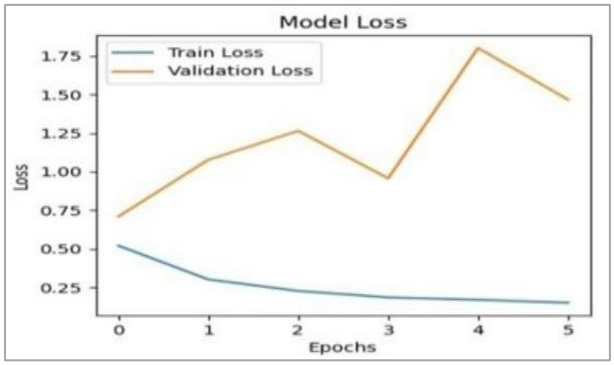

Fig. 8. Training and validation loss over 5 epochs (short run).

Figure 8 (short run) illustrates validation loss rising after the second epoch, suggesting overfitting when training is truncated.

Figure 9 reveals rapid training accuracy growth (**>90%**) while validation accuracy remains low (**≈45%**), further indicating overfitting in short runs.

Figure 10 (extended training) demonstrates improved generalization: training accuracy reaches **≈98%**, while validation accuracy rises to **≈92%**, showing stable performance with longer training and appropriate regularization.

Overall, extended training with data augmentation (compression, noise, resolution scaling) and early stopping/learning-rate scheduling improved generalization and reduced overfitting.

CONCLUSION

This paper presented a multi-modal deepfake detection framework that fuses visual and audio analysis to detect manipulated media. The visual pipeline combines CNN backbones (ResNet, EfficientNet) with LSTM-based temporal modelling to capture spatial and temporal inconsistencies, while the audio pipeline extracts spectral descriptors (MFCCs, mel-spectrograms) processed by a 2D CNN and a Random Forest classifier. An attention-based late-fusion module aggregates modality embeddings and outputs calibrated confidence scores.

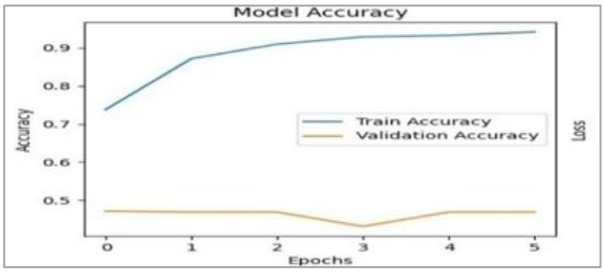

Fig. 9. Training and validation accuracy over 5 epochs (short run).

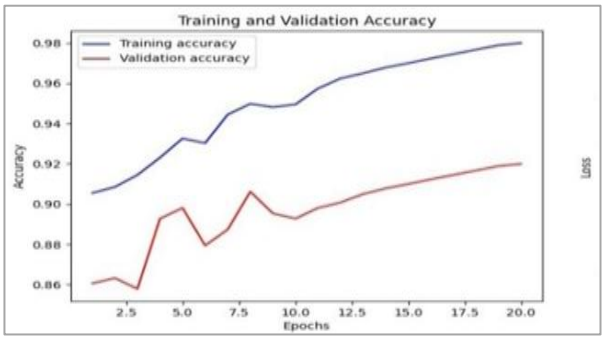

Fig. 10. Training and validation accuracy over 20 epochs (extended run).

Experimental evaluation across images, videos, and audio samples shows that the proposed system reliably detects deepfakes and provides interpretable outputs (bounding boxes and confidence values). Training and validation analyses indicate that extended training with augmentation substantially improves generalization (training **≈98%**, validation **≈92%** in representative

***Corresponding Author:** Nandana K Gowda

runs). Future work will expand dataset diversity, improve robustness to unseen forgery types, and optimize the system for large-scale, low-latency deployment.

FUTURE WORK

This project establishes a foundation for multi-modal deepfake detection by combining image, video, and audio analysis. Future research can build upon this framework by improving adversarial robustness to withstand increasingly sophisticated deepfake generation methods and by optimizing the models for real-time performance on mobile and edge devices. Expanding the system to include additional modalities, such as text analysis, physiological signals, or multilingual speech, can further enhance detection accuracy. Researchers may also explore explainable AI techniques to generate interpretable outputs that aid forensic experts and judicial authorities in evidence verification. Furthermore, the development of more diverse, large-scale datasets and privacy-preserving training methods, such as federated learning, will enable greater generalization and secure deployment. Ultimately, future work should focus on creating scalable, transparent, and legally reliable systems that integrate seamlessly into real-world forensic and media authentication pipelines.

ACKNOWLEDGEMENT

The authors confirm that there are no conflicts of interest, financial or personal, that could have influenced or are relevant to the content and outcomes of this research. We also express our gratitude to the contributors of the open datasets used in this study.

REFERENCES

1. Kim, E., & Cho, S. (2021). Exposing fake faces through deep neural networks combining content and trace feature extractors. *IEEE Access, 9*, 123493–123503. https://doi.org/10.1109/ACCESS.2021.3110859
2. Kang, J., Ji, S.-K., Lee, S., Jang, D., & Hou, J.-U. (2022). Detection enhancement for various deepfake types based on residual noise and manipulation traces. *IEEE Access, 10*, 69031–69047. https://doi.org/10.1109/ACCESS.2022.3185121
3. Tran, V.-N., Kwon, S.-G., Lee, S.-H., Le, H.-S., & Kwon, K.-R. (2023). Generalization of forgery detection with meta deepfake detection model. *IEEE Access, 11*, 535–550. https://doi.org/10.1109/ACCESS.2022.3232290
4. Khalifa, A. H., Zaher, N. A., Abdallah, A. S., & Fakhr, M. W. (2022). Convolutional neural network based on diverse Gabor filters for deepfake recognition. *IEEE Access, 10*, 22678–22680. https://doi.org/10.1109/ACCESS.2022.3152029
5. Guo, H., Hu, S., Wang, X., Chang, M.-C., & Lyu, S. (2022). Robust attentive deep neural network for detecting GAN-generated faces. *IEEE Access, 10*, 32574–32577. https://doi.org/10.1109/ACCESS.2022.3157297
6. Avilés-Cruz, C., & Celis-Escudero, G. J. (2023). 3G-AN: Triple-generative adversarial network under coarse-medium-fine generator architecture. *IEEE Access, 11*, 105344–105350. https://doi.org/10.1109/ACCESS.2023.3317897
7. Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense CNN architecture for deepfake image detection. *IEEE Access, 11*, 22081–22099. https://doi.org/10.1109/ACCESS.2023.3251417
8. Alnaim, N. M., Almutairi, Z. M., Alsuwat, M. S., Alalawi, H. H., Alshobaili, A., & Alenezi, F. S. (2023). DFFMD: A deepfake face mask dataset for infectious disease era with deepfake detection algorithms. *IEEE Access, 11*, 16711–16714. https://doi.org/10.1109/ACCESS.2023.3246661
9. Yue, P., Chen, B., & Fu, Z. (2024). Local region frequency guided dynamic inconsistency network for deepfake video detection. *IEEE Access, 7*(3), 889–904. https://doi.org/10.26599/ACCESS.2024.9020030
10. Khan, S. A., & Dang-Nguyen, D.-T. (2024). Deepfake detection: Analyzing model generalization across architectures, datasets, and pre-training paradigms. *IEEE Access, 12*, 1880–1885. https://doi.org/10.1109/ACCESS.2023.3348450
11. Liu, Q., Xue, Z., Liu, H., & Liu, J. (2024). Enhancing deepfake detection with diversified self-blending images and residuals. *IEEE Access, 12*, 46109–46114. https://doi.org/10.1109/ACCESS.2024.3382196
12. Khormali, A., & Yuan, J.-S. (2024). Self-supervised graph transformer for deepfake detection. *IEEE Access, 12*, 58114–58118. https://doi.org/10.1109/ACCESS.2024.3392512
13. Park, J., Park, L. H., Ahn, H. E., & Kwon, T. (2024). Coexistence of deepfake defenses: Addressing the poisoning challenge. *IEEE Access, 12*, 11674–11678. https://doi.org/10.1109/ACCESS.2024.3353785
14. Gautam, V., Kaur, G., Malik, M., et al. (2024). FFDL: Feature fusion-based deep learning method utilizing federated learning for forged face detection. *IEEE Access, 13*, 5366–5371. https://doi.org/10.1109/ACCESS.2024.3523257
15. Hydara, E., Kikuchi, M., & Ozono, T. (2024). Empirical assessment of deepfake detection: Advancing judicial evidence verification through artificial intelligence. *IEEE Access, 12*, 151188–151196. https://doi.org/10.1109/ACCESS.2024.3480320
16. Raza, S. A., Habib, U., Usman, M., Cheema, A. A., & Khan, M. S. (2024). MMGANGuard: A robust approach for detecting fake images generated by GANs using multi-model techniques. *IEEE Access, 12*, 104153–104156. https://doi.org/10.1109/ACCESS.2024.3393842
17. Safwat, S., Mahmoud, A., Fattoh, I. E., & Ali, F. (2024). Hybrid deep learning model based on GAN and ResNet for detecting fake faces. *IEEE Access,*

***Corresponding Author:** Nandana K Gowda

12, 86391–86405. https://doi.org/10.1109/ACCESS.2024.3416910

18. Felouat, H., Nguyen, H. H., Le, T.-N., Yamagishi, J., & Echizen, I. (2024). eKYC-DF: A large-scale deepfake dataset for developing and evaluating eKYC systems. *IEEE Access, 12*, 30876–30882. https://doi.org/10.1109/ACCESS.2024.3369187
19. Alrowais, F., Hassan, A. A., Almukadi, W. S., Alanazi, M. H., Marzouk, R., & Mahmud, A. (2024). Boosting deep feature fusion-based detection model for fake faces generated by generative adversarial networks for consumer space environment. *IEEE Access, 12*, 147680–147683. https://doi.org/10.1109/ACCESS.2024.3470128
20. Tchaptchet, E., Tagne, E. F., Acosta, J., Rawat, D. B., & Kamhoua, C. (2024). Deepfakes detection by iris analysis. *IEEE Access, 13*, 8977–8980. https://doi.org/10.1109/ACCESS.2024.3527868

***Corresponding Author:** Nandana K Gowda