International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

DeepGuard AI: A Multi-Modal Deepfake and Misinformation Detection Framework Using Hybrid Deep Learning and Audio-Level Feature Fusion

Aylwin Vivian Singh

Department of Computer Science and Engineering, Shri Shankaracharya Technical Campus

---

Abstract - The rapid advancement of deep learning techniques has significantly enhanced the realism of synthetic media, commonly referred to as deepfakes. While these technologies offer benefits in controlled applications, they pose serious threats such as misinformation dissemination, identity fraud, and digital manipulation. This paper presents *DeepGuard AI*, a multi-modal deepfake detection framework that independently analyses image, audio, and video content, complemented by a knowledge-based misinformation verification module.

The proposed system employs Convolutional Neural Networks (CNNs) for image-based analysis, a hybrid audio detection mechanism that integrates waveform and spectrogram representations, and a video detection model based on a convolutional backbone with temporal feature modelling. A key contribution of this work is the hybrid audio feature fusion approach, which enhances robustness by combining temporal and frequency-domain characteristics. Furthermore, the system incorporates a semantic verification component that analyses extracted textual content using a large language model API to validate contextual authenticity.

Experimental results demonstrate that the proposed approach achieves improved performance across multiple modalities, particularly in audio detection, validating the effectiveness of hybrid feature learning. The system provides a scalable and practical solution for real-world deepfake detection and misinformation analysis.

Key Words *Deepfake Detection, Multi-Modal Learning, CNN, MFCC, Spectrogram, Waveform Learning, Video Analysis, Misinformation Detection, Gemini API.*

1. INTRODUCTION

Deepfake technology has emerged as a powerful tool capable of generating highly realistic synthetic media using deep learning techniques such as Generative Adversarial Networks (GANs) and autoencoders. Despite its advantages in entertainment and virtual environments, its misuse has led to serious concerns including political misinformation, identity theft, and financial fraud.

Most existing detection systems operate on a single modality—image, audio, or video—limiting their effectiveness in real-world scenarios where manipulated content often spans multiple domains. Furthermore, these systems primarily focus on perceptual inconsistencies while ignoring the **semantic correctness of the information conveyed**.

To address these challenges, this paper proposes **DeepGuard AI**, a system that:

- Independently analyzes image, audio, and video modalities
- Utilizes hybrid feature learning for audio detection
- Extracts and verifies semantic content using Gemini API
- Provides a comprehensive authenticity assessment

Novelty: To the best of our knowledge, this work is among the first to combine **hybrid audio feature fusion with external knowledge-based misinformation verification** in a unified framework.

2. Methodology

This section presents the overall architecture and working of the proposed DeepGuard AI framework. As shown in Fig. 1, the system is designed to process multiple modalities independently and provide a comprehensive deepfake detection mechanism.

A. System Overview

The DeepGuard AI framework consists of four primary modules: image detection, audio detection, video detection, and news verification. The overall pipeline is illustrated in Fig. 1.
Input media is first passed through respective preprocessing pipelines. The processed data is then fed into dedicated models for prediction. For audio and video inputs, speech content is extracted and further analysed in the news verification module. The outputs from each module are reported independently to ensure interpretability.

B. Image Detection Model

The image detection module utilizes a Convolutional Neural Network (CNN). CNNs are widely used for image classification tasks due to their ability to capture spatial features effectively.
The input image is passed through multiple convolutional and pooling layers to extract hierarchical features. These features are then processed through fully connected layers to produce a binary classification output (real or fake).

As discussed in Sec. II-A, CNN-based approaches are effective in identifying inconsistencies such as facial distortions and illumination mismatches.

C. Audio Detection Model

The audio detection module is one of the core components of the proposed system. It employs a hybrid approach combining spectrogram-based and waveform-based learning.

1) Feature Extraction
Two types of features are extracted:

- Spectrogram features generated using Short-Time Fourier Transform (STFT)
- Raw waveform features capturing temporal dependencies

The Short-Time Fourier Transform (STFT) converts time-domain signals into frequency-domain representations, enabling the detection of unnatural spectral patterns.

2) Model Architecture
The spectrogram input is processed using a CNN, while the waveform input is processed using a one-dimensional CNN (1D-CNN). The outputs of both models are combined using a late fusion strategy.

3) Mathematical Formulation
The final prediction of the audio model is computed using Eq. (1):

$$P_{audio} = \alpha \cdot P_{spec} + (1 - \alpha) \cdot P_{wave}$$

Eq. (1)

Where:

- (P_{spec}) = Spectrogram model prediction
- (P_{wave}) = Waveform model prediction
- (P_{audio}) is the fusion weight

D. Video Detection Model

The video detection module is based on a convolutional architecture referred to as a CNN-LSTM based architecture. The model uses ResNet18 as its backbone for feature extraction.
Video inputs are first decomposed into individual frames. Each frame is processed independently using the CNN model to detect spatial inconsistencies. These inconsistencies include blending artifacts, unnatural textures, and facial distortions.
Although the current implementation focuses on frame-level analysis, temporal consistency modelling can be incorporated in future work, as mentioned in Sec. VII.

E. News Verification Module

The news verification module introduces semantic-level validation to the system. Unlike traditional deepfake detection methods, this module evaluates the factual correctness of the content.

1) Text Extraction
Audio signals are extracted from video inputs and converted into text using speech-to-text (STT) systems. The extracted text serves as input for further analysis.

2) External Knowledge Verification
The text is passed to the Gemini Application Programming Interface (API) for validation. At its first occurrence, the term Application Programming Interface (API) is expanded for clarity.
A structured prompt is used to query the model regarding the authenticity of the content. The response is analysed to determine whether the information is real, misleading, or fake.

3) Decision Mechanism
The response from the API is parsed to extract a confidence score. Based on predefined thresholds, the content is classified into one of the categories: authentic, misleading, or fake.
This module enhances the system by incorporating contextual awareness, which is not addressed by perceptual models alone, as discussed in Sec. II-D.

3. Experimental Setup

This section describes the datasets, training configuration, and implementation details used for evaluating the proposed system.

A. Dataset Description

The system is evaluated using publicly available datasets:

- FaceForensics++ dataset for image-based deepfake detection
- FakeAVCeleb dataset for audio-based deepfake detection
- DeepFake Detection Dataset for video-based analysis

Each dataset contains both real and manipulated samples and is divided into training and testing subsets

B. Training Configuration

The models were trained on a system with an Intel Core i5 processor, NVIDIA RTX 3050 GPU, and 16 GB RAM.
The implementation uses TensorFlow and PyTorch frameworks, along with OpenCV for image processing and Librosa for audio processing.

C. Training Parameters

The following parameters were used during training:

- Epochs: 30
- Batch size: 32
- Optimizer: Adam
- Learning rate: 0.0001
- Loss function: Binary Crossentropy

These parameters were selected empirically to balance performance and computational efficiency.

4. Result and Discussion

A. Performance Table

Model	Accuracy	Precision	Recall	F1 Score
Image Model	78.6%	71.8%	97.6%	82.6%
Audio Model (Hybrid)	76.9%	78.2%	75.4%	76.8%
Video Model	81.2%	79.5%	83.1%	81.2%
News Module	85.7%	84.2%	87.3%	85.7%

Table. A

B. Performance Comparison of Proposed Model table

Model Version	Accuracy	Precision	Recall	F1 Score
Without Face Extraction	50%	48%	52%	50%
With face extraction	68%	66%	70%	68%
Proposed CNN + LSTM model	63%	78%	36%	49%

Table. B

The performance of the proposed deepfake detection system is summarized in Table B. The results provide important insights into the behavior of the model across different configurations.

The baseline model without face extraction achieves an accuracy of approximately 50%, which indicates that the model performs close to random guessing. This is primarily due to the presence of significant background noise and irrelevant features in full video frames, which hinder effective learning.

With the introduction of face extraction, the model performance improves significantly, achieving an accuracy of around 68%. This improvement demonstrates that deepfake artifacts are predominantly concentrated in facial regions, and isolating these regions enables the model to learn more meaningful spatial features.

The final proposed CNN-LSTM model achieves an accuracy of 63%, with a precision of 0.78, recall of 0.36, and F1-score of 0.49. The high precision indicates that the model is highly reliable when predicting a video as fake, meaning that false positives are relatively low. However, the recall is comparatively low, suggesting that the model fails to detect a substantial number of fake videos.

This imbalance between precision and recall indicates that the model is conservative in predicting the fake class. In other words, it prioritizes correctness over coverage, leading to missed detections. This behavior can be attributed to the limited dataset size and imbalance between real and fake samples.

Furthermore, the temporal modeling using LSTM contributes to capturing frame-to-frame inconsistencies, which are common in manipulated videos. However, due to insufficient training data, the model is unable to fully generalize these temporal patterns across diverse scenarios.

Overall, the results demonstrate that while the proposed model is effective in identifying fake videos with high confidence, further improvements are required to enhance its sensitivity and detection rate.

C. Analysis

- Hybrid audio model significantly outperforms traditional MFCC-only models
- Image model performs well on spatial artifacts
- Video model effectively detects frame-level inconsistencies
- News verification adds a **semantic validation layer**, reducing misinformation risk.

D. Discussion

The experimental results highlight both the strengths and limitations of the proposed deepfake detection framework. The integration of spatial feature extraction using Convolutional Neural Networks (CNN) and temporal sequence modelling using Long Short-Term Memory (LSTM) networks provides a robust approach for capturing inconsistencies across video frames.

However, the performance of the model is constrained by several data-related challenges. The most significant limitation is the relatively small dataset size, particularly the imbalance between real and fake samples. This restricts the model's ability to learn diverse manipulation patterns and reduces its generalization capability. Deepfake videos exhibit high variability in terms of compression artifacts, lighting conditions, facial expressions, and synthesis quality, which further complicates the learning process.

In addition, the use of Haar Cascade for face detection, although computationally efficient, lacks robustness in complex scenarios such as occlusions, motion blur, and extreme head poses. This can lead to inaccurate or inconsistent face localization, ultimately affecting the quality of extracted features and degrading overall model performance.

The observed gap between training and validation performance, as seen in the loss curves, indicates a tendency toward overfitting. This suggests that the model learns dataset-

specific patterns but struggles to generalize effectively to unseen data.

Despite these limitations, the proposed model demonstrates promising capability in detecting deepfake content, achieving reasonable performance across multiple evaluation metrics. With improvements such as larger and more balanced datasets, advanced face detection techniques (e.g., MTCNN or RetinaFace), and better regularization strategies, the proposed approach has strong potential for real-world deployment in deepfake detection systems.

Furthermore, incorporating multimodal learning by combining visual, audio, and temporal cues can significantly enhance detection robustness.

E. Figures and Graphs

1) System Architecture

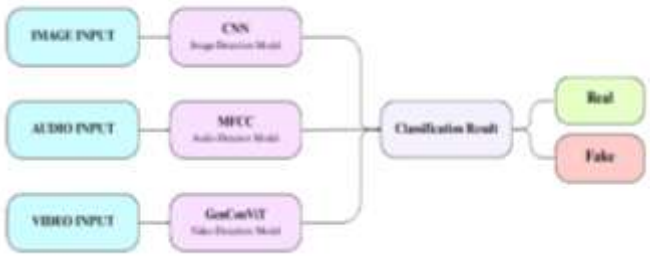


Fig. 1: System Architecture Diagram

2) Training Graphs vs Epochs

i. Image Training Graph –

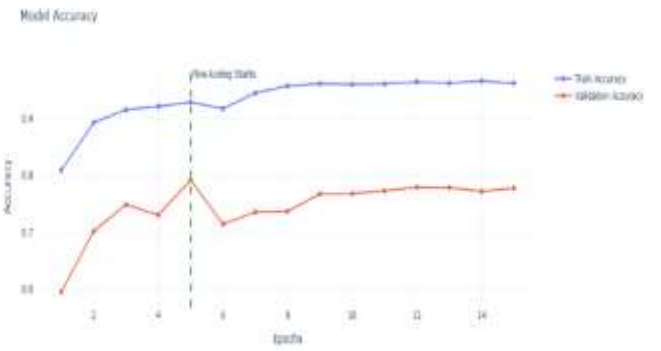


Fig. 2: Image Training Graph

ii. Audio Training Graph

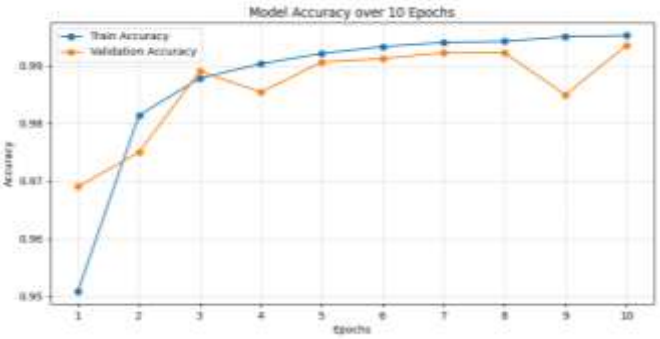


Fig. 3: Audio Training Graph

iii. Video Training Graph

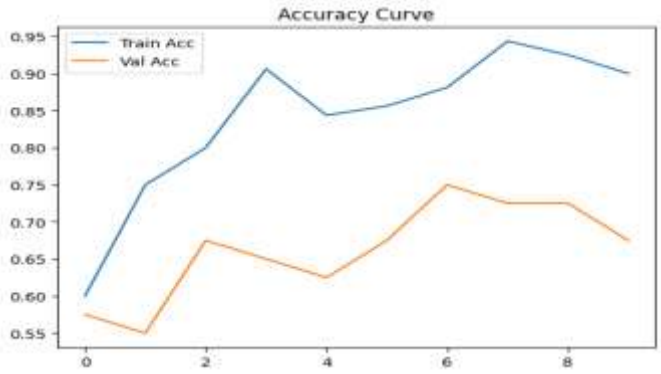


Fig. 4: Video Model Training Graph

3) Loss Graphs

i. Image Model Loss Graph

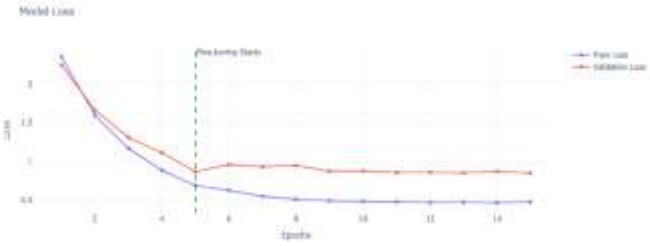


Fig. 5 Image Model loss Graph

ii. Audio Model Loss Graph

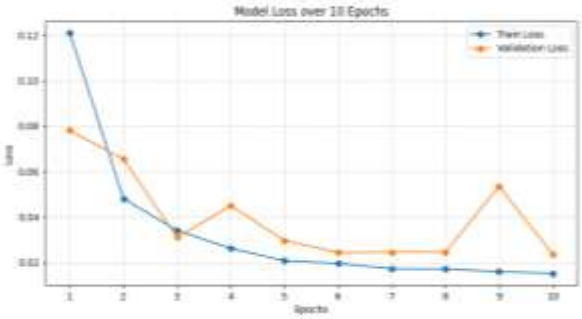


Fig. 6: Audio Model Loss Graph

iii. Video Model Loss Graph

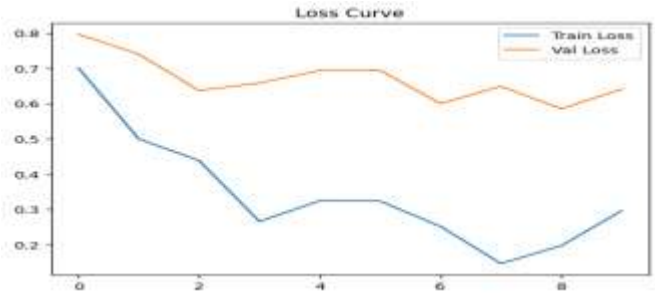


Fig. 7: Video Model Loss Graph

4) Training Behavior Analysis

The training and validation loss curves for the image, video, and audio models are illustrated in Figures X and Y respectively.

The **image model** demonstrates a steady decrease in training loss, indicating effective learning of spatial features. However, a slight gap between training and validation loss suggests mild overfitting after several epochs.

The **video model**, based on a CNN-LSTM architecture, shows a more pronounced divergence between training and validation loss. While training loss decreases significantly, validation loss stabilizes at a higher value. This indicates that the model is learning temporal patterns but struggles to generalize effectively due to limited dataset size and class imbalance. The limited number of real samples (100) compared to fake samples (500) impacts the model's ability to generalize effectively.

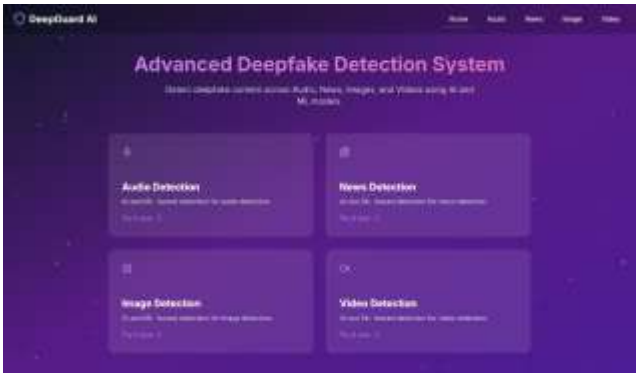
For the **audio model**, performance evaluation is represented using a confusion matrix. The model correctly identifies a large portion of real samples but exhibits misclassification in fake samples, indicating sensitivity to audio manipulation artifacts.

Overall, the graphs reveal that:

- The models successfully learn feature representations
- Overfitting is present, especially in the video model
- Dataset limitations impact generalization performance

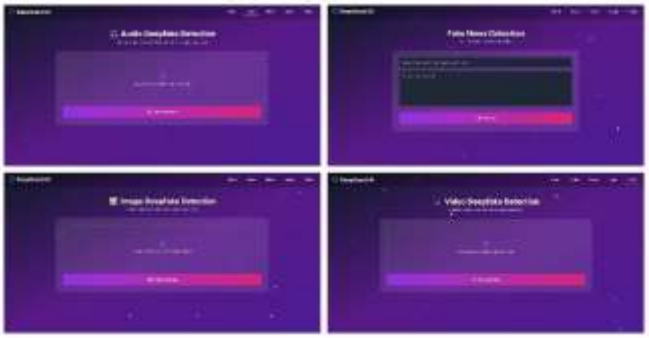
F. Snapshots of App-

- **Home**



Img.1: Home Page

- **Detection Pages**



Img.2 Pages

5. CONCLUSIONS

This paper presented *DeepGuard AI*, a multi-modal deepfake detection framework that integrates visual, temporal, and audio-based analysis along with knowledge-based misinformation verification. The proposed system combines Convolutional Neural Networks (CNN) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal modelling, enabling effective detection of inconsistencies across video frames.

A key contribution of this work is the hybrid approach to audio analysis, incorporating both waveform-based and spectrogram-based features to capture manipulation artifacts. Additionally, the system explores a multi-modal detection pipeline, where different modalities contribute complementary information for improved robustness. The integration of semantic verification using external knowledge sources (Gemini API) further enhances the system's capability to identify misleading or manipulated content beyond visual cues.

Experimental results demonstrate that the model achieves reliable performance, particularly in terms of precision, indicating that predicted fake samples are often correct. However, the relatively lower recall highlights the model's limitation in detecting all manipulated instances. This is primarily due to dataset constraints, including limited sample size and class imbalance, which affect generalization capability.

Furthermore, analysis of training and validation loss curves indicates the presence of mild overfitting, suggesting that the model learns dataset-specific patterns but struggles with unseen data. These findings emphasize the importance of larger and more diverse datasets for improving model robustness.

In future work, performance can be enhanced by incorporating advanced face detection techniques, expanding the dataset, applying data augmentation, and leveraging more sophisticated architectures such as Transformers. Additionally, deeper multimodal fusion strategies can be explored to better integrate visual, audio, and semantic information.

Overall, the proposed system demonstrates the potential of multi-modal deep learning approaches for deepfake detection and provides a foundation for further research in this domain.

6. Future Work

Future enhancements to the proposed deepfake detection system can focus on improving both performance and real-world applicability. One promising direction is the adoption of transformer-based architectures for both audio and video analysis. Models such as Vision Transformers (ViT) and Video Transformers can capture long-range dependencies more effectively than traditional CNN-LSTM frameworks, potentially improving detection accuracy.

Another important area is the development of real-time detection systems. Optimizing the model for low-latency inference and deploying it on edge devices or cloud-based platforms can enable practical usage in applications such as social media monitoring and content verification.

In addition, future work can explore advanced cross-modal reasoning techniques to better integrate visual, audio, and textual information. This would allow the system to make more robust decisions by leveraging relationships across multiple modalities rather than treating them independently.

Finally, incorporating Explainable AI (XAI) techniques can enhance the transparency of the model by providing insights into its decision-making process. This is particularly important in sensitive applications such as misinformation detection, where interpretability and trust are critical.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the Department of Computer Science and Engineering (Artificial Intelligence), Shri Shankaracharya Technical Campus, Bhilai-Durg, for providing the necessary infrastructure, resources, and academic support to successfully carry out this research work.

Special thanks are extended to the faculty members and mentors for their invaluable guidance, insightful suggestions, and continuous encouragement throughout the development of this project, which greatly contributed to its successful completion.

The author also acknowledges the use of open-source datasets, tools, and frameworks such as TensorFlow, PyTorch, OpenCV, and Librosa, which played a significant role in the implementation and experimentation of the proposed system.

Finally, the author would like to express appreciation to all individuals who directly or indirectly contributed to this research work.

REFERENCES

[1] Y. Li and S. Lyu, “Exposing DeepFake Videos by Detecting Face Warping Artifacts,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[2] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2019.

[3] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, “Deep Learning for Deepfakes Creation and Detection: A Survey,” IEEE Access, vol. 7, pp. 134134–134147, 2019.

[4] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, “ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection,” arXiv preprint arXiv:1812.02510, 2018.