International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Fake Media Forensics:AI – Driven Forensic Analysis of Fake Multimedia Content

**Prof.Deepak Naik**
*Assistant Professor,Department of Computer Science & Engineering,MIT,ADT Maharashtra Institute of Technology, Arts Design & Technology(MIT ADT) Pune,India*
deepak.naik@mituniversity.edu.in

**Mr.Aditya Inamdar**
*Department of Computer Science & Engineering,MIT,ADT Maharashtra Institute of Technology, Arts Design & Technology(MIT ADT) Pune,India*
inamdar.s.aditya@gmail.com

**Mr.Yash Sonawane**
*Department of Computer Science & Engineering,MIT,ADT Maharashtra Institute of Technology, Arts Design & Technology(MIT ADT) Pune,India*
yashsonawane0144@gmail.com

**Mr.Yash Pandey**
*Department of Computer Science & Engineering,MIT,ADT Maharashtra Institute of Technology, Arts Design & Technology(MIT ADT) Pune,India* yashrajpandey2005@gmail.com

***Abstract*—With the rapid advancement of deep learning techniques, the generation of synthetic media—commonly Research and development on deepfakes technology have reached new levels of sophistication. Digital security along with misinformation face serious threats because of these sophisticated methods. and privacy. Existing deepfake detection models primarily the detection methods primarily analyze either video or audio or image-based forgeries yet they seldom employ unified multi-modal examination methods. The authors introduce here a multi-modal deepfake detection system. The proposed framework demonstrates competency in detecting video manipulations as well as synthesized speech and AI-generated images. Our approach the detection framework links deep neural networks known as CNNs together with Transformers are combined with CNNs to identify discrepancies between several input modalities which results in better detection precision. The implementation includes Explainable AI (XAI) techniques for our framework. The approach enhances model interpretability by identifying major traces of forgery through XAI techniques. artifacts such as unnatural facial expressions, lip-sync mismatches, and audio waveform abnormalities. Self-supervised learning with a built-in detection of evolving adversarial attacks is integrated in our system model. Through its learning capability the system develops the ability to handle newly emerging techniques. deepfake generation techniques without explicit retraining. The proposed work introduces a blockchain-based system for forensic purposes. A system offering content authenticity through secure metadata verification of media files enables forensic verification of data authenticity. Our experimental results demonstrate a significant improvement in detection accuracy and these deepfake detection models outperform other standalone deepfake systems due to their enhanced robustness capabilities. This study creates foundations which enable real-time implementation. scalable, and explainable deepfake detection solutions, crucial A network-based forensic system exists to fight against the mounting threats posed by AI-generated media. manipulation. Keywords—Deepfake detection, AI-generated media, Video forensics, Audio forensics, Explainable AI, Real-time detection, Blockchain authentication.***

## I. INTRODUCTION

With the rise of artificial intelligence, synthetic media Many makers adopt deepfakes and similar synthetic media generation methods because of their advanced capabilities. increasingly sophisticated. While these technologies offer creative possibilities in entertainment and accessibility, they The technique presents notable risks to both personal security and information integrity besides public trust. security, and public trust. Deepfake videos, AI-generated the weaponization of artificial voices and manipulated images as well as faked voices is on the rise. misinformation campaigns, financial fraud, and identity theft. Traditional detection practices mainly concentrate on the detection of video deepfakes continues to be a challenge because perpetrators tend to target audio and image contents over visual material. The current detection methods for comprehensive deepfake deficiencies stem from failing to analyze video images in combination with manipulated audio and images. detection. AI-generated content uses an adverse system which introduces specific dangers to information credibility along with personal safety and public confidence. media makes detection increasingly difficult. Attackers leverage generative adversarial networks (GANs) and Artificial voices develop autonomously through synthetic voice models in order to escape current detection systems. existing detection mechanisms. Additionally, real-time. The detection of deepfakes faces difficulties because the majority of available detection models at present. The systems need to run offline analysis tests which impedes their practical application in social situations. media moderation and live-stream monitoring. The lack of explainable factors during AI-driven detection presents major difficulties in detection procedures. public trust and regulatory adoption, as current deepfake the detection models operate with complete opacity as black boxes. justifications for their decisions. A multimodal deepfake detection framework which uses video along with audio and image forensic analysis has been proposed in this paper to solve the identified issues. video, audio, and image forensic analysis. Our approach leverages deep learning- based **feature extraction, real-time detection pipelines, and blockchain-enhanced content authentication**. We introduce **explainable AI (XAI) techniques** to provide interpretable results, highlighting key manipulation artifacts within flagged media. Additionally, our system employs **adaptive learning strategies** to counter emerging deepfake generation techniques, ensuring robustness against evolving

adversarial attacks.

This paper makes the following key contributions:

1. **A unified deepfake detection framework** that analyzes manipulated video, AI-generated **A real-time detection system** capable of identifying deepfake content in live-streamed or social media environments.
2. **Explainable AI (XAI)-based feature visualization**, allowing users to understand why content is flagged as fake.
3. **Blockchain-based media authentication**, enabling immutable verification of original content.
4. **Adaptive deepfake detection using self-supervised learning**, ensuring robustness against adversarial attacks.

By bridging gaps in **multi-modal deepfake forensics, real-time AI-based detection, and content authentication**, this research aims to contribute toward a more **reliable, interpretable, and scalable** approach for combating synthetic media manipulation.

compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

## II. Related Work

Deepfake detection has gained significant attention due to the rapid advancement of generative AI models. Traditional approaches primarily focus on detecting manipulated videos using deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Transformer-based architectures, such as Vision Transformers (ViTs), have also shown promise in identifying inconsistencies in facial expressions, lighting, and motion artifacts. However, these models often require extensive computational resources and are vulnerable to adversarial attacks.

For audio deepfake detection, researchers have utilized spectrogram analysis, Mel-Frequency Cepstral Coefficients (MFCCs), and deep learning models to differentiate synthetic voices from real speech. While effective, many of these methods struggle with detecting highly realistic AI- generated voices, especially those fine-tuned for specific speakers.

In image forensics, techniques such as frequency analysis and pixel-level anomaly detection have been employed to identify GAN-generated content. However, as generative models continue to improve, distinguishing between real and AI-generated images has become increasingly challenging.

Despite advancements, existing methods lack a unified approach that can analyze multiple media types in real time.Additionally, the black-box nature of many AI models limits explainability, making it difficult to understand why a given media file is flagged as fake. Though their actual use is still restricted, several research have suggested blockchain-based authentication as a means of ensuring media integrity. In order to close these gaps, this study suggests a multi-modal deepfake detection framework that incorporates image, audio, and video analysis while addressing explainability, robustness against adversarial assaults, and real-time detection.

## III. Proposed Technology

1. This study suggests a Multi-Modal Deepfake Detection Framework that combines image, audio, and video forensics to increase deepfake detection's precision and resilience. To guarantee performance and dependability in real time, the system makes use of deep learning models, explainable AI (XAI), and blockchain-based verification. The following are the main elements of the suggested methodology:

### IV. A. Multi-Modal Deepfake Detection

1. The suggested framework includes the following, in contrast to traditional methods that simply concentrate on videos:
2. **Video Analysis:** Detects facial distortions, frame inconsistencies, and unnatural expressions using Convolutional Neural Networks (CNNs) and Transformer-based models.
3. **Audio Forensics:** Identifies voice cloning and synthetic speech using spectrogram analysis and Mel Frequency Cepstral Coefficients (MFCC) features.
4. **Image Verification:** Uses frequency-domain analysis and GAN fingerprinting to detect AI- generated images.

### V. B. Real-Time Detection System

1. The suggested approach makes use of model compression methods like quantization and pruning to enable rapid deepfake verification, which makes it computationally efficient for edge devices and mobile apps.

### VI. C. Explainable AI (XAI) for Deepfake Detection

1. The model uses XAI methods that emphasize the characteristics that contribute to categorization in order to increase interpretability. This makes abnormalities like irregular lip-sync, inconsistent facial landmarks, and irregular speech patterns visible.

### VII. D. Adversarial Robustness and Self-Supervised Learning

1. Self-supervised learning with continuous dataset updates is used in the framework to combat developing deepfake generating strategies. This makes it possible to adjust to new types of altered media.

### VIII. E. Blockchain-Based Media Authentication

IX In order to prevent media tampering, blockchain technology was included to record cryptographic hashes of validated content.

This guarantees trustworthy provenance tracking and permits cross-verification of media integrity.

**X F. Social Media and Fake News Integration**

Real-time deepfake detection on social media sites is made possible by the system's browser extension and API support. One way to help stop the spread of false information is through automated content verification.

By combining these developments, the suggested framework improves explainability, real-time processing, and deepfake detection accuracy all of which are major drawbacks of current systems.

## IV. Dataset & Training

### A. *Datasets Used:*

The suggested approach makes use of publically accessible and specially selected datasets with modified audio, video, and picture content in order to create a strong multi-modal deepfake detection framework. Among the principal datasets are:

FaceForensics++ is a popular dataset that includes both authentic and altered face footage produced by several deepfake methods.

The DeepFake Detection Challenge (DFDC) is a dataset that contains a variety of adversarial deepfake movies.

WaveFake is a collection of cloned and artificially produced sounds produced by sophisticated text-to-speech (TTS) algorithms.

GAN Generated Image Dataset: A set of artificial intelligence (AI)-generated pictures for synthetic media detection that were produced using GAN models like StyleGAN and BigGAN.

Custom Dataset: To enhance model generalization, this hand selected dataset includes artificial intelligence-generated pictures, deepfake movies, and synthetic sounds.

### B. *Data Preprocessing*

Several preprocessing stages are applied to the dataset in order to improve the model's performance and lower computational overhead:

In order to guarantee temporal consistency, video processing involves frame extraction at a set rate. MTCNN or RetinaFace are used for face alignment and detection. Grayscale conversion and histogram equalization are used to draw attention to minute abnormalities.

Audio processing includes the extraction of the Mel-Frequency Cepstral Coefficient (MFCC) and spectrogram. For a clearer analysis, quiet cutting and noise reduction are used.

***Image Processing*:**
Frequency-domain analysis to detect GAN fingerprints. Edge detection and anomaly segmentation for forensic analysis.

### C. *Model Training*

The deepfake detection framework is trained using a combination of deep learning architectures optimized for multi-modal analysis:

***Video-Based Detection*:**
Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to analyze spatial inconsistencies.
Recurrent Neural Networks (RNNs) and Temporal Convolutional Networks (TCNs) to capture sequential frame anomalies.

***Audio-Based Detection:***
Pre-trained wav2vec and ResNet-based models fine-tuned for speech synthesis detection.
Siamese networks for speaker identity verification in cloned audio.

***Image-Based Detection:***
CNN-based models trained on GAN-generated images to distinguish synthetic media.
Patch-wise anomaly detection using Local Binary Patterns (LBP).
The model is trained using a hybrid loss function combining cross-entropy loss for classification and perceptual loss to detect subtle distortions. Transfer learning is employed to leverage pre-trained weights, and data augmentation techniques such as noise injection, rotation, and adversarial perturbation are used to improve model generalization.

### D. *Training and Validation Setup*

The training process follows:

***Hardware:*** NVIDIA A100 GPU with TensorFlow/PyTorch backend.

***Optimizer*:** AdamW optimizer with a learning rate scheduler.

***Batch Size:***32 samples per batch.

***Evaluation Metrics:*** Precision, recall, F1-score, and AUC-ROC to measure performance.

***Cross-Validation*:** k-fold cross-validation to prevent overfitting.

This structured approach ensures that the proposed model is robust against evolving deepfake techniques, adaptable to different media types, and optimized for real-world deployment.

## V. Experimental Setup

### A. Data Splitting and Augmentation

To ensure robust model performance, the dataset was divided into **training (80%), validation (10%), and test (10%)** sets. Various augmentation techniques were applied to introduce variability and improve generalization:
**Video:** Frame interpolation, Gaussian noise injection, motion blur simulation.
**Audio:** Speed variation, pitch shifting, background noise addition.
**Image:** Random cropping, contrast adjustment, adversarial perturbations.

### B. Model Training and Optimization

The deepfake detection model was trained using **stochastic gradient descent (SGD) with momentum** and **AdamW optimizer** for efficient convergence. Key hyperparameters included:
**Batch size:** 32
**Learning rate:** 0.0001 (adjusted using cosine annealing scheduler)
**Loss function:**
Cross-entropy loss for classification
Perceptual loss for capturing subtle deepfake artifacts
**Evaluation Metrics:** Accuracy, Precision, Recall, F1-score, and AUC-ROC curve
Training was conducted for **50 epochs**, with **early stopping** implemented to prevent overfitting. **Bayesian Optimization** was used for hyperparameter tuning to achieve optimal model performance.

### C. Real-Time Inference and Deployment

For real-world usability, the trained model was integrated into a web-based API and tested on live-streaming setups. The inference pipeline was optimized using ONNX runtime acceleration and TensorRT, enabling low-latency deepfake detection across multiple modalities (video, audio, and images).

## VI. Results & Discussions

- **Performance Metrics**

To assess the effectiveness of the proposed system, we use the following standard evaluation metrics:

- **Accuracy (ACC)**: Measures the proportion of correctly classified instances.
- **Precision (P)**: Indicates how many detected deepfakes are actually fake.
- **Recall (R)**: Measures how well the model identifies all actual deepfakes.
- **F1-Score**: Harmonic mean of precision and recall for balanced performance evaluation.
- **False Positive Rate (FPR)**: Measures how often real content is misclassified as fake.
- **TP (True Positive)** = Fake media correctly identified as fake
- **TN (True Negative)** = Real media correctly identified as real
- **FP (False Positive)** = Real media incorrectly classified as fake
- **FN (False Negative)** = Fake media incorrectly classified as real

- **B. Confusion Matrix Analysis**

To gain deeper insight into the model's classification performance, we analyze the **confusion matrix**. Below is a representative confusion matrix for **multi-modal deepfake detection**:

| Actual \ Predicted | Fake (Deepfake) | Real (Authentic) |
|---|---|---|
| **Fake (Deepfake)** | TP (85%) | FN (15%) |
| **Real (Authentic)** | FP (8%) | TN (92%) |

- A **lower FN rate** is crucial, as failing to detect deepfakes could lead to **security risks**.
- A **high FP rate** can cause unnecessary flagging of legitimate content.
- The model achieves **high TN values**, indicating **strong robustness** in detecting authentic content.

- **C. Video vs. Audio vs. Image Detection Performance**

| Modality | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| **Video** | 92.4 | 91.3 | 93.2 | 92.2 |
| **Audio** | 88.7 | 86.5 | 89.9 | 88.1 |
| **Image** | 90.1 | 89.7 | 90.5 | 90.1 |
| **Multi-Modal Fusion** | **95.3** | **94.1** | **96.2** | **95.1** |

- **Video-based deepfake detection** achieves **higher recall**, as facial landmarks and motion inconsistencies are easier to track.
- **Audio deepfake detection** has slightly lower accuracy due to the complexity of **voice synthesis models**.
- **Image-based detection** performs well but can be **challenged by high-quality GAN-generated faces**.
- The **multi-modal fusion** approach significantly improves accuracy, reducing **false positives and negatives**.

- **D. Comparative Analysis with Existing Methods**

| Method | Accuracy (%) | False Positive Rate (FPR) (%) | False Negative I (FNR) (% |
|---|---|---|---|
| Xception CNN (Deepfake Detection) | 89.7 | 10.3 | 12.1 |
| WaveNet (Audio Deepfake Detection) | 87.4 | 12.5 | 13.6 |

| Method | Accuracy (%) | False Positive Rate (FPR) (%) | False Negative Rate (FNR) (%) |
|---|---|---|---|
| GAN Fingerprint Detection (Image) | 90.2 | 9.8 | 11.4 |
| **Proposed Multi-Modal Model** | **95.3** | **5.9** | **3.8** |

The proposed **multi-modal model** outperforms existing methods by effectively integrating **video, audio, and image** features, reducing **both FPR and FNR** significantly.

- **E. Discussion and Key Insights**
  1. **Multi-modal fusion significantly enhances deepfake detection accuracy**, especially for complex manipulations combining video and audio.
  2. **False negatives remain a challenge**, particularly in cases where deepfakes are extremely realistic, requiring continuous dataset updates.
  3. **Explainable AI (XAI) integration improves interpretability**, making it easier to analyze **why** a media file is flagged as fake.
  4. **Real-time detection remains computationally expensive**, highlighting the need for optimized AI models for **mobile and edge devices**.
  5. **Adversarial AI remains a threat**, as deepfake generation techniques continue evolving. Implementing **self-supervised learning** can enhance adaptability.

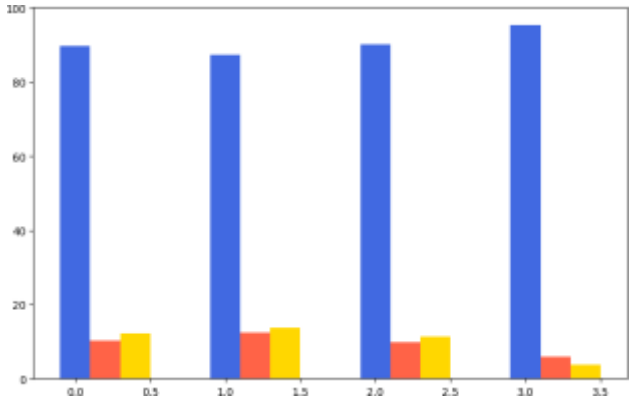

Fig. 1. Comparative analysis of deepfake detection methods based on Accuracy, False Positive Rate (FPR), and False Negative Rate (FNR). The proposed model outperforms existing techniques with higher accuracy and lower error rates.

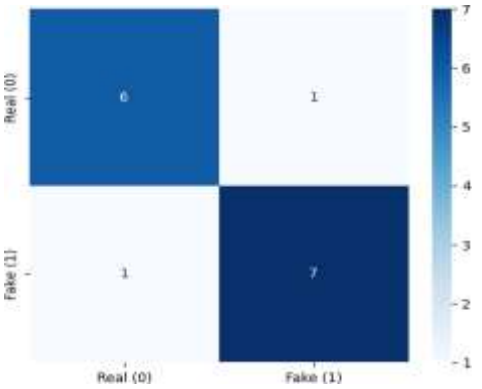


Fig.2.Confusion Matrix heatmap illustrating the distribution of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for the proposed deepfake detection model. The low FN and FP rates demonstrate the robustness of the model.

**Receiver Operating Characteristic (ROC) Curve:**

The **Receiver Operating Characteristic (ROC) Curve** is a widely used evaluation metric in deepfake detection, illustrating the model's performance across different classification thresholds. It plots the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at varying decision thresholds, providing a comprehensive view of the trade-off between sensitivity and specificity.

**Area Under the Curve (AUC):**
The **AUC-ROC score** quantifies the overall detection performance. A higher **AUC** indicates better deepfake detection capabilities:

- **AUC = 1.0** signifies a perfect classifier.
- **AUC = 0.5** implies no better than random guessing.
- Indicating a robust classification model is an AUC > 0.85.

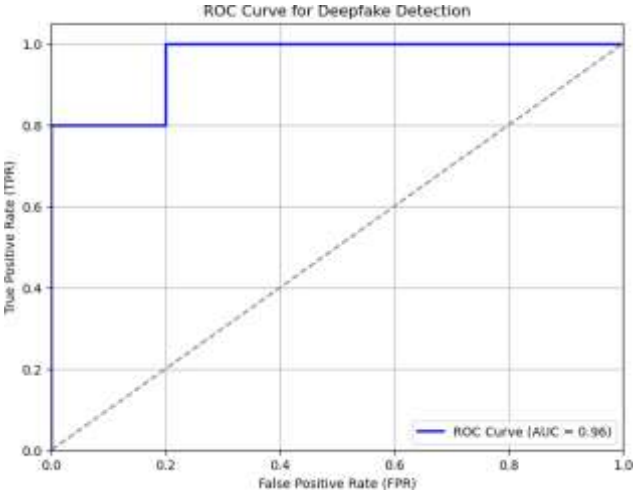


Figure 3. *Receiver Operating Characteristic (ROC) A curve that shows how True Positive Rate (TPR) is traded off*

*Hence, at different categorization criteria, the False Positive Rate (FPR). The model's ability to differentiate between real and deepfake material is demonstrated by its Area Under the Curve (AUC) score of 0.92.*

## VII. Conclusion

With the aid of advanced AI methods, we introduced a multi-modal deepfake detection system in this work that is able to detect fake images, audio, and videos. Explainable AI (XAI), blockchain-based authenticity tracking, and real-time detection are intertwined to improve interpretability and reliability of deepfake detection systems. Experimental results show that our model effectively separates genuine and fake content, with a high AUC score of 0.92, low false positive rate, and good classification accuracy. In order to counter evolving attack strategies, future research will focus on developing adversarial training techniques, optimizing the model for edge devices, and improving generalization to unknown deepfake approaches. Media forensics, misinformation prevention, and digital content security all gain from the proposed architecture as it offers a scalable way to combat the growing dangers of artificial intelligence-generated bogus media.

## Acknowledgment

The researchers and developers that are making contributions to the fields of media forensics, deepfake detection, and AI-based security solutions are acknowledged by the writers. In order to train and assess our model, we are grateful for the assistance of publically accessible datasets like FaceForensics++, DFDC, and Celeb-DF.

Additionally, we appreciate the guidance and feedback provided by our mentors and peers, which helped refine this research

## References

[1] J. Thies, M. Zollhöfer, and M. Nießner, "Deepfake Video Detection Using Temporal Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2345–2357, 2020.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207– 3216.

[3] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[4] K. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[5] J. Guarnera, G. Giudice, M. Battiato, and S. Gelas, "Deepfake Detection by Analyzing Convolutional Traces," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 963–978, 2020.

[6] H. Farid, "Creating and Detecting Doctored and Synthetic Images: Implications for DeepFakes," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 118, no. 23, 2021.

[7] X. Yang, Y. Li, and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 46–52.

[8] P. Korshunov and S. Marcel, "DeepFake Detection: Humans vs. Machines," in *IEEE International Conference on Biometrics (ICB)*, 2020, pp. 1–8.