TRUSTLENS: A MULTI-MODAL DEEPFAKE DETECTION FRAMEWORK FOR DIGITAL MEDIA FORENSICS

Name of Author: Yash Rajesh Kelvalkar
MSc Information and Cyber Security Student
Department of Information Technology
Guru Nanak Khalsa College, Mumbai, India
Email: g24.yash.kelvalkar@gnkhalsa.edu.in

Abstract

The rapid advancement of artificial intelligence has significantly improved the realism of synthetic media generation. While such developments offer innovation in digital communication and content creation, they also introduce serious risks related to misinformation, fraud, and document forgery. Deepfakes synthetically generated or manipulated images, videos, audio, text, and documents have become increasingly difficult to distinguish from authentic content. This raises an urgent need for practical and reliable verification mechanisms.

This paper presents **TrustLens**, a unified multi-modal deepfake detection framework designed to authenticate five categories of digital media: images, videos, audio recordings, textual content, and PDF documents. The system integrates an EfficientNet-B0-based image classifier, a custom convolutional neural network trained on MFCC features for audio analysis, frame-based video verification, a RoBERTa-based text classification module, and digital signature verification techniques for PDF documents.
Unlike single-modality detection systems, TrustLens combines independent verification mechanisms into a modular and CPU-compatible architecture deployed through a Streamlit web interface. Experimental results demonstrate strong performance across modalities while maintaining computational efficiency suitable for practical deployment. The proposed framework contributes toward scalable digital forensic systems capable of addressing emerging multi-modal deepfake threats.

Keywords:

Deepfake detection, multi-modal analysis, Efficient Net, Mel-frequency cepstral coefficients features, Digital forensics, Media authentication, Digital signature verification, AI-generated content

1. Introduction

Artificial intelligence has transformed digital content creation over the past decade. Among its most disruptive developments is deepfake technology, which enables the generation of highly realistic synthetic media. With publicly accessible tools and increasing computational accessibility, producing manipulated visual, audio, and textual content has become significantly easier.

Although deepfake technology has constructive applications in entertainment, accessibility, and creative industries, its misuse poses serious threats. Incidents involving voice-cloned financial fraud, AI-generated misinformation, and forged digital documents illustrate the potential societal impact of synthetic media. As generative models continue to evolve, the challenge of distinguishing authentic from manipulated content becomes increasingly complex.

Most existing detection approaches focus on a single data modality, such as image classification or audio analysis. However, contemporary manipulations frequently span multiple media types. A forged video may include synthetic audio and AI-generated captions, while tampered documents may embed manipulated images. Addressing these challenges requires an integrated verification strategy rather than isolated detection pipelines.

This research introduces TrustLens, a unified framework that combines five detection mechanisms into a single deployable architecture. The goal is not only to achieve strong detection performance but also to ensure practical usability in real-world digital forensic contexts.

1.1 Research Contribution

The primary contribution of this work lies in the integration of five independent deepfake detection and verification mechanisms into a unified, modular, and CPU-compatible system. Unlike many prior studies that focus solely on audio-visual fusion or single-modality classification, TrustLens incorporates image, video, audio, text, and PDF verification within a single architecture.

Additionally, the inclusion of encrypted document handling and digital signature validation extends the framework beyond conventional deepfake detection toward applied digital forensics. The modular design enables scalability and future enhancement without restructuring the system.

2. Literature Review

Deepfake detection research has evolved alongside generative modeling techniques. Early detection methods concentrated primarily on visual artifacts introduced by generative adversarial networks (GANs). Convolutional neural networks demonstrated effectiveness in identifying inconsistencies in texture, facial geometry, and frequency-domain characteristics.

Subsequent research emphasized multi-modal detection strategies, suggesting that combining audio and visual features improves robustness. Fusion approaches early, intermediate, and late fusion have been explored to enhance cross-modal learning. However, such systems often remain limited to audio-visual integration and do not extend to textual or document-level verification.

EfficientNet architectures have gained recognition for their balanced trade-off between computational efficiency and classification accuracy. EfficientNet-B0, in particular, provides strong performance with relatively low parameter complexity, making it suitable for CPU-based deployment.

Audio deepfake detection commonly employs MFCC (Mel-Frequency Cepstral Coefficient) features to capture spectral patterns associated with synthetic speech. Convolutional neural networks trained on MFCC representations have shown promising results in identifying cloned or text-to-speech audio.

Transformer-based models such as RoBERTa have demonstrated strong contextual understanding capabilities for distinguishing human-written and AI-generated text. Their use in authenticity classification continues to expand.

Document forensic research focuses on metadata analysis, structural validation, and digital signature verification. As digitally signed PDFs become standard in government and enterprise workflows, automated verification tools are increasingly relevant.

Despite progress in individual domains, comprehensive frameworks integrating visual, audio, textual, and document verification remain limited. TrustLens addresses this gap by offering a unified detection environment.

3. Research Methodology

3.1 System Architecture

TrustLens employs a modular architecture where each media type is processed independently before results are presented through a unified interface. This design ensures flexibility and simplifies model updates.

The system consists of:

- Image Detection Module: EfficientNet-B0 CNN
- Video Detection Module: Frame extraction combined with image classifier
- Audio Detection Module: Custom CNN trained on MFCC features
- Text Detection Module: RoBERTa-based classifier
- PDF Verification Module: Digital signature and structural analysis

The modular approach enables scalability while maintaining computational practicality.

3.2 Dataset Description

Modality	Dataset Details	Characteristics
Image	Source: Multi-domain deepfake collections Classes: Real (authentic) and Fake (GAN-generated, face-swapped)	Size: 635 images Split: 381 train, 127 val, 127 test Format: JPEG, PNG (various resolutions)
Audio	Source: Generated audio samples with filename-based labeling Classes: Real (human voice) and Fake (synthetic, TTS)	Size: Variable clips (3-sec processed) Features: 16kHz sampling, 40 MFCC coefficients
Video	Source: Video clips analyzed through frame extraction Classes: Inherited from image classification	Processing: 8 keyframes per video Format: MP4, AVI, MOV
Text	Source: Human-written and AI-generated samples Classes: Human-written and AI-generated (GPT-based)	Processing: TF-IDF vectorization / transformer input Pre-trained models used
PDF	Source: Tecxoft samples, government documents Verification: AcroForm, SigFlags, encryption metadata	Classes: Signed (valid digital signatures) and Unsigned/Tampered

Table 1: Dataset overview across five modalities

3.3 Implementation Details

Framework: PyTorch 2.x for deep learning models

Libraries:

- torchvision: Image transformations and pre-trained models
- librosa: Audio processing and MFCC extraction
- PyPDF2: PDF structure analysis
- scikit-learn/joblib: Text model serialization
- OpenCV: Video frame extraction
- Streamlit: Web interface deployment

Deployment Platform: Streamlit Cloud / Local server

Hardware Requirements: CPU-compatible (tested on Windows 11, Python 3.12)

3.4 Evaluation Metrics

Primary Metrics:

- **Accuracy:** Percentage of correctly classified samples
- **Precision:** True positives / (True positives + False positives)
- **Recall:** True positives / (True positives + False negatives)
- **F1-Score:** Harmonic mean of precision and recall

Confidence Scoring:

- Probability-based confidence (0-100%)
- Classification thresholds: FAKE (>70%), SUSPICIOUS (50-70%), REAL (<50%)

4. Data Analysis

The performance evaluation of TrustLens across five modalities demonstrates strong detection capabilities, with results summarized in the following table:

Modality	Architecture	Dataset Size	Accuracy	Inference Time
Image	EfficientNet-B0	635 images (381 train, 127 val)	94.5%	<1 second
Audio	Custom CNN (MFCC features)	Variable clips (3-sec processed)	100%*	1-2 seconds
Video	Frame-based (Image model)	8 frames/video	$\sim$94%**	3-5 seconds
Text	RoBERTa-based Classifier	Pre-trained (external data)	$>$90%	<1 second
PDF	Digital Signature Verification	Signed/Unsigned (Tecxoft + Govt)	$\sim$100%	<1 second

Table 2: Performance summary across TrustLens modalities (*validation set accuracy; **frame-level estimate)

The results indicate strong performance across all modalities. It is important to note that reported metrics reflect validation dataset performance. Broader cross-domain evaluation would provide deeper insight into generalization capability.

5. Practical and Ethical Implications

Systems such as TrustLens may support government agencies in validating digitally signed documents, assist organizations in preventing fraud, and contribute to content moderation efforts. However, responsible deployment is critical. False positives may have reputational consequences, and uploaded media should be processed with appropriate privacy safeguards. Continuous monitoring and transparent governance mechanisms are necessary for ethical application.

6. Limitations

Several limitations should be acknowledged. The dataset size for certain modalities, particularly image and audio training data, was relatively limited. The video module relies on frame-based analysis rather than explicit temporal modeling, which may reduce sensitivity to subtle motion artifacts. Additionally, adversarial robustness testing against intentionally evasive deepfake techniques was not conducted.
Future large-scale benchmarking and cross-dataset evaluation will be necessary to strengthen robustness validation.

7. Conclusion

This study introduced TrustLens, a multi-modal deepfake detection framework integrating image, video, audio, text, and document verification within a unified architecture. Experimental evaluation demonstrated strong performance while maintaining CPU compatibility and practical deployability.

As synthetic media generation continues to evolve, adaptable and multi-layered verification systems will be essential. TrustLens provides a foundation for scalable and extensible digital forensic applications.

8. Future Study

Future enhancements may include:

- Integrating LSTM/GRU and 3D CNN architectures for improved temporal modeling
- Expanding datasets from large-scale benchmarks
- Implementing explainable AI visualizations such as Grad-CAM
- Optimizing inference through model quantization and GPU acceleration
- Exploring cross-modal fusion strategies
- Investigating blockchain-based content provenance
- Developing public APIs and browser extensions

These improvements can further strengthen TrustLens as an adaptive defense mechanism against evolving synthetic media threats.

References

[1] Researchers (2024). A Multimodal Framework for Deepfake Detection. *arXiv preprint arXiv:2410.03487*. https://arxiv.org/abs/2410.03487

[2] Authors (2023). A Robust Approach to Multimodal Deepfake Detection. *PMC Journal*, 10299653. https://pmc.ncbi.nlm.nih.gov/articles/PMC10299653/

[3] Authors (2024). A survey of digital forensic methods for multimodal deepfake detection. *PMC Journal*, 11157519. https://pmc.ncbi.nlm.nih.gov/articles/PMC11157519/

[4] Author (2024). Deep Fake Detection with EfficientNet. *Kaggle*. https://www.kaggle.com/code/kameshrasu/deep-fake-detection-with-efficientnet

[5] Authors (2025). Deepfake Detection using Efficientnet-B0 and GRU. *IJERT*. https://www.ijert.org/deepfake-detection-using-efficientnet-b0-and-gru

[6] Authors (2025). Dual-Branch Fusion Model for Deepfake Detection. *PMC Journal*, 12295270. https://pmc.ncbi.nlm.nih.gov/articles/PMC12295270/

[7] Authors (2025). Deepfake audio detection via MFCC features and mel-spectrogram using deep learning. *AIP Conference Proceedings*, 3264(1), 030027. https://pubs.aip.org/aip/acp/article/3264/1/030027/3338488

[8] Authors (2025). Audio Deepfake Detection Using MFCC-SVM, CQCC-GMM. *IJIREEICE*. https://ijireeice.com/wp-content/uploads/2025/11/IJIREEICE.2025.131103-audio.pdf

[9] Authors (2024). RoBERTa-BiLSTM Approach to Detect AI-Generated Text. *arXiv preprint arXiv:2407.02978*. https://arxiv.org/html/2407.02978v1

[10] Authors. RoBERTa and Bi-LSTM for Human vs AI Generated Text. *CEUR Workshop Proceedings*, Vol-3740, paper-272. https://ceur-ws.org/Vol-3740/paper-272.pdf

[11] Authors (2022). Forensic authenticity examination of PDF documents. *SPIE Digital Library*, 12506, 2662214. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12506/2662214

[12] Authors (2025). Cyber Forensic and Reverse Engineering Techniques for Digital Signature Verification. *IJPSAT*. https://ijpsat.org/index.php/ijpsat/article/view/6976