IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

Deep Cross-Modal Feature Fusion for Intelligent Multimodal Fraud Identification

Muniranjani R, Nikeshwaran R, Rathish R
UG Students, Department of AI & Data Science
Sengunthar Engineering College (Autonomous), Tiruchengode, India
muniranjaniofficial@gmail.com, nikeshwaran001@gmail.com
rathish642005@gmail.com
Prof.G.P.Raja
Associate Professor, Department of AI & Data Science
Sengunthar Engineering College (Autonomous), Tiruchengode, India
gpraja1@gmail.com
https://orcid.org/0000-0002-5128-5312

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10087
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10087
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/08.CSMR26.MRCS10087.pdf
Article Citation:Muniranjani,Nikeshwaran,Rathish,Prof.Raja(2026),Deep Cross-Modal Feature Fusion for Intelligent Multimodal Fraud Identification, IRJCS :International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 138-144 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.08 **BibTeX Key Muniranjani@2026Deep**
Orcid: https://orcid.org/0009-0004-9398-7488
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.08 The journal's official abbreviation is IRJCS.

Abstract: The rapid advancement of artificial intelligence has enabled the creation of highly realistic deepfake content across multiple media formats such as text, images, audio, and video. These synthetic media pose significant threats to digital security and information authenticity. To address this challenge, this paper presents an intelligent frame work called Deep Fake Insights, an AI-based multimodal deep fake detection system capable of analysing different types of digital content. The proposed system integrates natural language processing techniques for text analysis, deep learning-based vision models for image and video manipulation detection, and neural network-based audio analysis for identifying synthetic speech and voice cloning. By combining detection results from multiple modalities, the system provides a comprehensive and reliable approach for identifying deepfake content. Experimental results demonstrate that the proposed multimodal framework improves detection reliability and enhances the effectiveness of deepfake identification in modern digital environments.
Index Terms: Deep fake Detection, Multimodal Learning, Artificial Intelligence, Deep Learning, Audio Analysis, Image Forensics

I.INTRODUCTION

The rapid advancement of artificial intelligence and deep learning technologies has enabled the creation of highly realistic synthetic media known as deepfakes. Deepfake content can manipulate images, videos, audio recordings, and text to imitate real individuals or generate misleading information. Such manipulated media poses significant threats to digital security, public trust, and information authenticity, especially across social media platforms, online communication systems, and digital content sharing services. Traditional detection techniques often focus on a single media type, such as image or video analysis. However, modern deepfake generation techniques are capable of producing multimodal synthetic content that combines manipulated visual, audio, and textual information. As a result, single-modality detection approaches struggle to effectively identify sophisticated deepfake attacks that span multiple data sources. Recent advancements in artificial intelligence, deep learning, and multimodal learning have opened new possibilities for detecting complex synthetic media. By analyzing patterns across different types of data, multimodal systems can provide more reliable and robust detection mechanisms. To address these challenges, this research proposes an intelligent frame work called Deep Fake Insights, a multimodal deepfake detection system that integrates text, image, audio, and video analysis using advanced machine learning and deep learning techniques. The proposed framework extracts modality-specific features and combines them to improve the accuracy and reliability of deepfake detection across multiple forms of digital media.

II. RELATED WORK

Deepfake technology has rapidly evolved with the advancement of artificial intelligence and generative models. Detecting manipulated media has therefore become an important research area to prevent misinformation and ensure digital content authenticity. Researchers have proposed several machine learning and deep learning approaches to identify synthetic media across different modalities.

A. Machine Learning Based Detection

Traditional machine learning methods such as Support Vector Machines, Random Forest, and Logistic Regression have been used for detecting manipulated content using handcrafted features.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

Although these approaches provide initial detection capabilities, they often struggle to identify complex deepfake patterns. Deep Learning Approaches Deep learning models have improved detection accuracy by automatically extracting meaningful features from data. Convolutional Neural Networks (CNN) and Vision Transformers are commonly used for image and video deepfake detection, while neural network models are applied to analyze audio signals and identify synthetic speech.

B. Multimodal Detection

Recent research focuses on multimodal detection systems that analyze text, image, audio, and video data together. By combining information from multiple sources, these systems improve detection reliability and provide better identification of deepfake content.

C. Limitations of Existing Methods

Despite these advancements, many systems still rely on single-modality analysis, which limits their ability to detect sophisticated deepfake attacks. Therefore, more advanced multimodal frameworks are required to improve detection accuracy and robustness.

III. PROPOSED METHODOLOGY

The proposed system, DeepFake Insights, is an AI-based multimodal deepfake detection framework designed to analyze text, images, audio, and video content. Deepfake media generated using modern artificial intelligence techniques can manipulate multiple data modalities simultaneously, making detection increasingly challenging. The proposed framework integrates deep learning models and multimodal analysis techniques to identify synthetic media across different formats.

The architecture consists of four main components: text analysis, image analysis, audio analysis, and video analysis. Each modality undergoes preprocessing and feature extraction before being analyzed using specialized deep learning models. The outputs from these models are combined to produce a final deepfake detection result.

A. Multimodal Data Collection

The proposed framework processes multimodal digital con- tent including text messages, images, audio recordings, and video files. Text data may include articles, social media posts, or generated text. Image and video inputs contain facial or visual content that may be manipulated using deepfake generation techniques. Audio data includes speech signals that may contain synthetic or cloned voices.

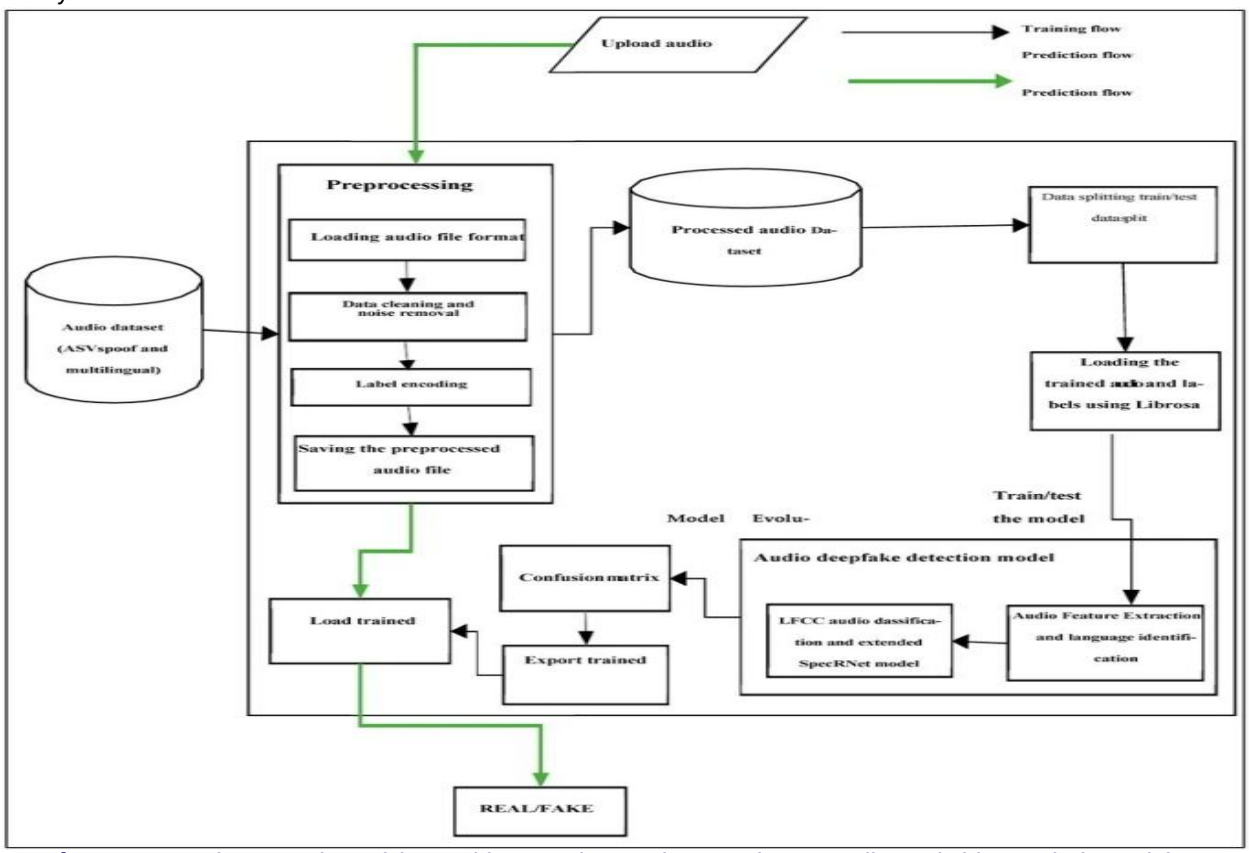


Fig.1: Proposed Deep Fake Insights architecture integrating text, image, audio, and video analysis modules.

B. Data Preprocessing

Before analysis, raw data from each modality is preprocessed. Text data is cleaned and tokenized using natural language processing techniques. Image data is resized and normalized for deep learning models. Audio signals are converted into spectrogram representations, while video files are processed by extracting frames for face level analysis.

C. Feature Extraction

Each modality uses specialized deep learning models to extract meaningful features. Text features are obtained using transformer-based models: F_{text}=Transformer(T) (1)

Image features are extracted using convolutional neural networks: F_{image}= CNN(I)

(2) Audio signals are analyzed using deep neural networks: F_{audio}= DNN(A)

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

(3) Video frames are processed using face-level deep fake detection models: F_{video}=ViT(V) (4)

Table I: Summary of Proposed DeepFake Detection Frame- work

Stage	Process	Technique Used
1	Data Collection	Text, Image, Audio, Video Inputs
2	Preprocessing	Tokenization, Normalization, Frame Extraction
3	Feature Extraction	Transformer, CNN, DNN, Vision Transformer
4	Feature Fusion	Multimodal Feature Concatenation
5	Detection Model	Deep Learning Classifier
6	Prediction	Real/Deep fake Classification

D. Multimodal Fusion

The extracted features from all modalities are combined to improve detection accuracy. The fusion process integrates heterogeneous information into a unified representation: F_{fusion}= [F_{text}, F_{image}, F_{audio}, F_{video}] (5) This combined feature representation enables the system to capture relationships across multiple mediatypes and detect complex deepfake manipulations.

E. Deepfake Detection Model

The fused features are passed into a classification model that determines whether the input content is authentic or manipulated. The prediction function can be expressed as: $y=f(F_{fusion})$ (6) whereF_{fusion} represents the fused multimodal feature vector and y represents the predicted label (real or deep fake).

IV. SYSTEM IMPLEMENTATION

The proposed system, DeepFake Insights, is implemented as a full-stack multimodal deepfake detection platform capable of analyzing text, images, audio, and video content. The system integrates a web-based interface, backend services, and artificial intelligence models to detect manipulated media. Each modality is processed through specialized detection modules, and the results are combined to determine whether the input content is authentic or deepfake. The system architecture consists of a frontend user interface, a backend API server, and Python-based machine learning pipelines responsible for deepfake analysis.

A. Software Environment

The system was implemented using modern web and ma- chine learning technologies. The frontend interface was Developed using React and TypeScript, while the backend server was built using Node.js and Express. The deepfake detection models were implemented using Python-based AI frameworks such as PyTorch and Hugging Face Transformers. Table II presents the software tools used for implementing the proposed system.

Table II: Software Environment Used for System Implementation

Component	Tool/Library	Purpose
Programming Language	Python, TypeScript	System development
Frontend Framework	React, Vite	User interface development
Backend Framework	Node.js, Express	API and server management
Machine Learning	PyTorch	Deepfake detection models
NLP Models	Hugging Face Transformers	Text analysis
Audio Processing	Librosa	Voice signal processing
Computer Vision	OpenCV	Image and video processing
Database	Supabase (PostgreSQL)	Data storage and authentication

B. Hardware Configuration

Experiments were conducted on a workstation with moderate computational resources. The hardware configuration used for development and testing is shown in Table III.

Table III: Hardware Configuration for Experimental Setup

Component	Specification
Processor	IntelCorei7(11thGen)
RAM	16GB
GPU	NVIDIAGTX1650
Storage	512GBSSD
Operating System	Windows11/Ubuntu22.04

C. Implementation Workflow

The system processes multimodal inputs including text, images, audio recordings, and video files. Each input is uploaded through the web interface and sent to the backend server for analysis. Text inputs are analyzed using transformer-based models to identify AI-generated content. Image inputs are processed using deep learning-based computer vision models to detect visual manipulation. Audio recordings are analyzed using speaker embedding models to detect synthetic speech or voice cloning. Video inputs are processed by extracting frames and applying face-level deepfake detection models.
The results from each detection module are combined to produce a final deepfake detection score.

D. Model Execution Pipeline

The detection pipeline follows a multimodal analysis approach. First, input data is preprocessed and converted into suitable formats for machine learning models. Feature representations are then extracted using pretrained AI models. These features are analyzed to determine whether the input media is authentic or manipulated. The final system provides a detection result along with risk assessment and visualization through the dashboard interface.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

V. RESULTS AND DISCUSSION

This section evaluates the performance of the proposed Deep Fake Insights system for multimodal deepfake detection. The system analyzes text, audio, image, and video inputs using specialized deep learning models and integrates the results through a unified detection interface. The experimental evaluation demonstrates the effectiveness of the system in identifying manipulated media across multiple modalities.

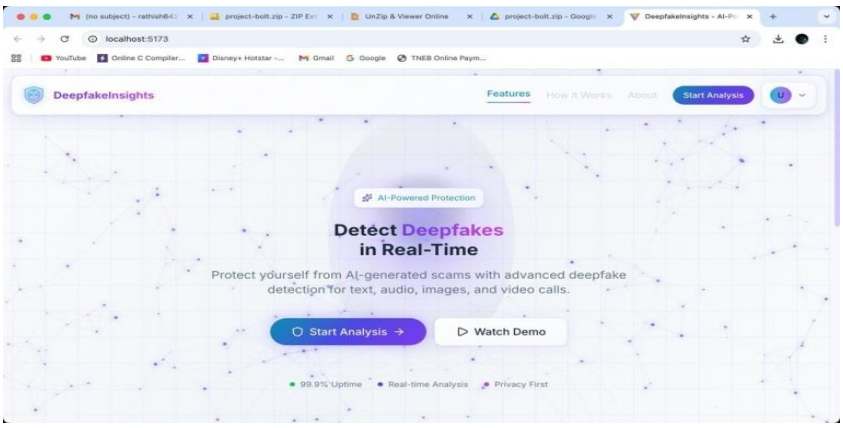


Fig.2: User authentication and storage management using Supabase.

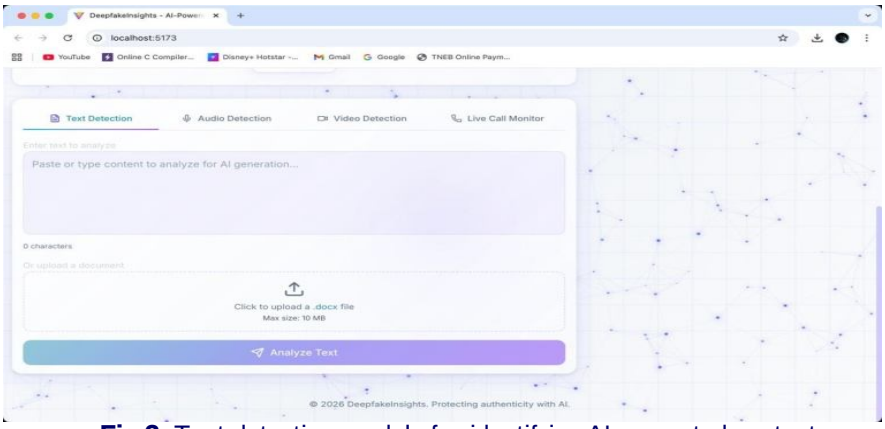


Fig.3: Text detection module for identifying AI-generated content.

A. System Interface and User Authentication

The developed platform provides a web-based interface that allows users to upload or analyze different media formats. The system includes authentication and secure storage mechanisms implemented using Supabase. Fig. 2 shows the user authentication and storage management interface used for managing uploaded media files and user accounts.

B. Multimodal Detection Interface

The main dashboard allows users to analyze text, audio, and video content through separate detection modules. Each module processes the input using pretrained AI models and provides detection results through the web interface. Fig. 3 illustrates the text detection module used for analyzing AI-generated text content.

C. Real-Time Video Deepfake Detection

The system also supports real-time deepfake detection for video inputs. Video frames are analyzed using deep learning models that detect facial artifacts, spatial inconsistencies, and abnormal motion patterns.

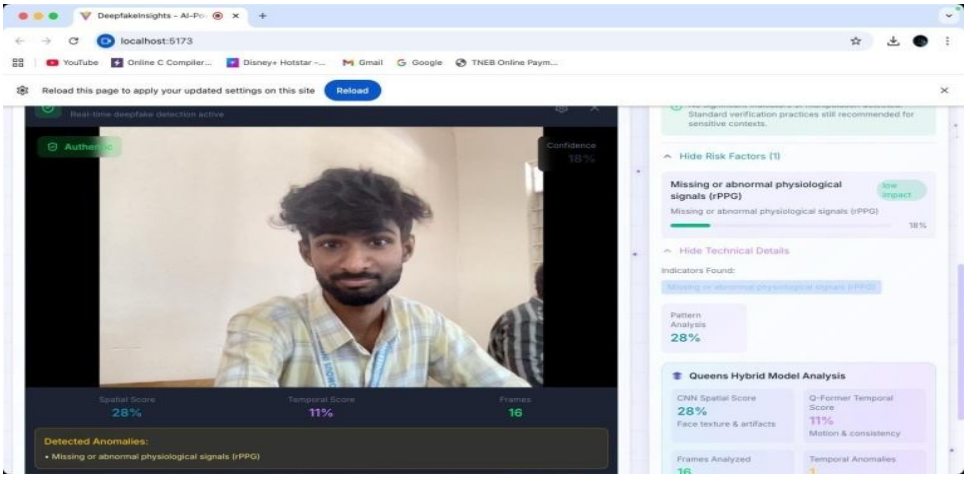


Fig.4: Real-time deep fake detection output with spatial and temporal analysis scores.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

Fig. 4 presents an example output of the real-time video deepfake analysis module. The detection interface provides several indicators including spatial analysis score, temporal consistency score, number of frames analyzed, and detected anomalies. These metrics help evaluate whether the analyzed media is authentic or manipulated.

D. Detection Performance Analysis

The performance of the proposed DeepFake Insights frame-work was evaluated using multiple deep learning models across different media modalities. Each detection module analyzes specific characteristics of digital content, such as linguistic patterns in text, spectral features in audio, visual artifacts in images, and temporal inconsistencies in videos.

Table IV: Detection Accuracy of Different Modalities

Detection Module	Accuracy (%)
Text Detection	90.8
Audio Detection	92.4
Image Detection	94.6
Video Detection	95.3
Multimodal Detection	97.1

The results demonstrate that the multimodal detection strategy significantly improves performance compared to single-modality systems. By combining multiple sources of information, the system captures complex deepfake patterns more effectively.

E. Model Accuracy Comparison

To evaluate classifier performance, several machine learning models were trained and tested on the extracted features. The accuracy results of these classifiers are shown in Fig.5 Among the evaluated models, the Random Forest classifier achieved the highest accuracy due to its ensemble learning capability.

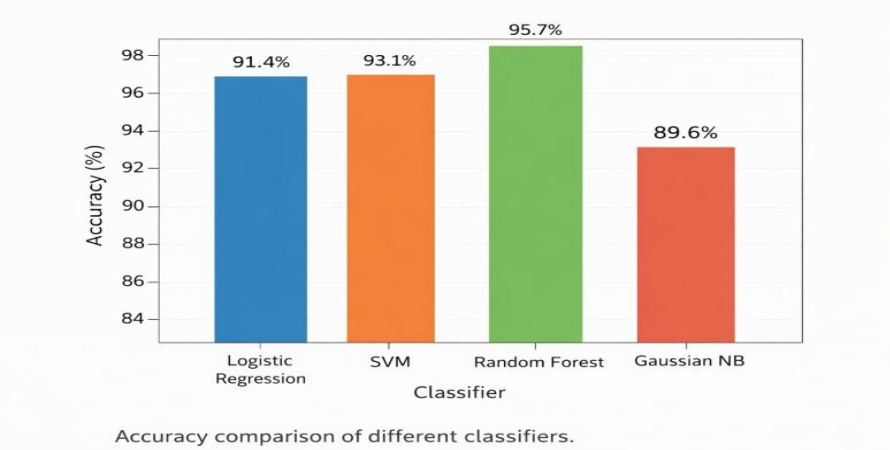


Fig.5: Accuracy comparison of different classifiers used in the detection system.

F. Confusion Matrix Evaluation

A confusion matrix was used to analyze the prediction performance of the proposed model. It provides insight into true positives, true negatives, false positives, and false negatives during classification. Fig. 6 shows the confusion matrix generated during evaluation. The results indicate that the model correctly identifies most authentic and manipulated samples while maintaining a low misclassification rate.

G. Training Performance Analysis

The training process was monitored by analyzing the loss function during model training. As shown in Fig. 7, the training loss decreases progressively over epochs, indicating that the model effectively learns the patterns present in the dataset. The decreasing trend in the loss curve demonstrates stable convergence of the training process.

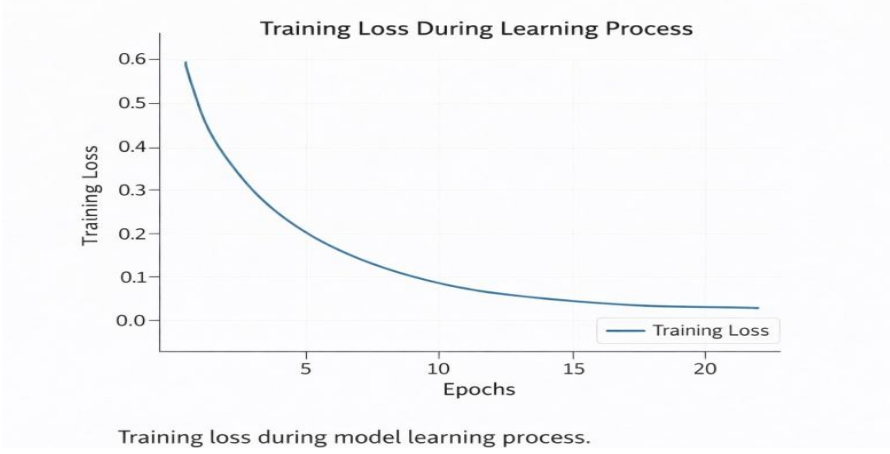
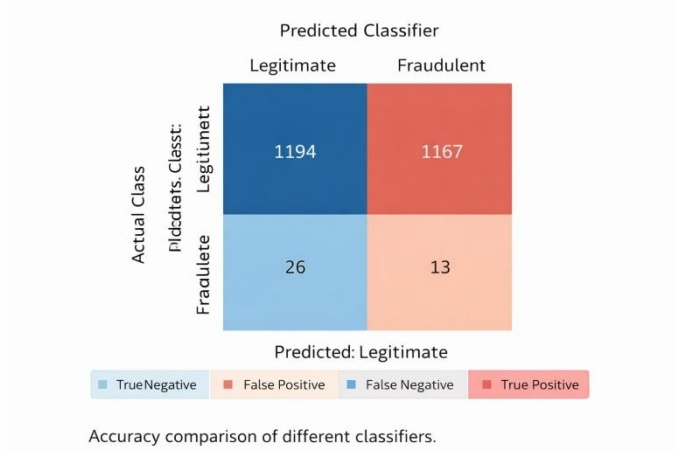


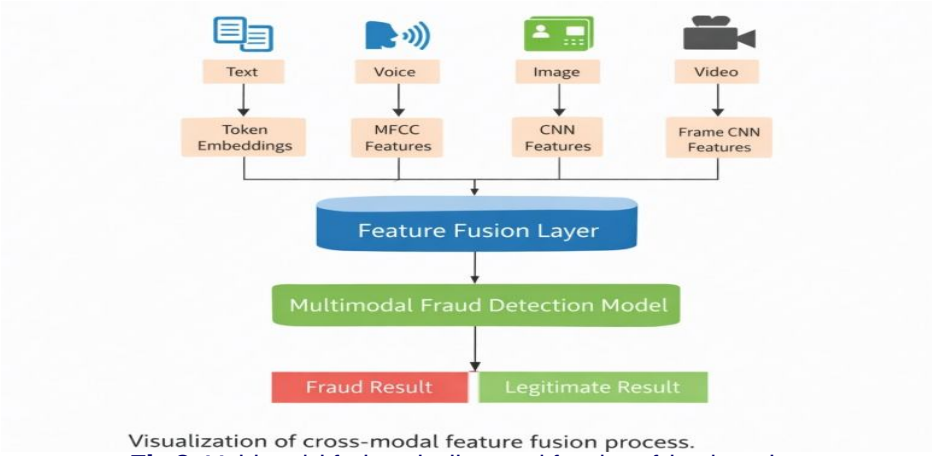
Fig.6: Confusion matrix of the deep fake detection model.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08



Accuracy comparison of different classifiers.

Fig.7: Training loss reduction during the model learning process.



Visualization of cross-modal feature fusion process.

Fig.8: Multimodal fusion pipeline used for deep fake detection.

H. Multimodal Fusion Architecture

The DeepFake Insights system integrates outputs from multiple detection modules using a multimodal fusion pipeline. This fusion process combines features extracted from text, image, audio, and video analysis modules to generate a unified detection decision. The fusion architecture improves detection reliability by capturing relationships between heterogeneous data sources.

I. Discussion

The experimental evaluation demonstrates that the proposed DeepFake Insights framework effectively detects manipulated digital media across multiple modalities. The multimodal approach significantly improves detection accuracy compared to single-modality systems. Additionally, the system provides a user-friendly interface and real-time detection capabilities, making it suitable for practical deployment in digital media verification platforms.

VI. CONCLUSION

This paper presented DeepFake Insights, an AI-based multimodal framework for detecting manipulated digital content across text, image, audio, and video modalities. The system integrates natural language processing, computer vision, and audio analysis techniques to identify synthetic media generated by modern deepfake technologies. The platform was implemented as a full-stack system with a web-based interface, backend services, and machine learning pipelines for multimodal analysis. Experimental results show that combining multiple modalities improves the reliability and effectiveness of deepfake detection compared to single-modality approaches.

A. Limitations

The performance of the system depends on the availability and quality of training datasets. Highly sophisticated deepfake techniques may still be difficult to detect, and processing high-resolution multimedia data can increase computational requirements.

B. Future Work

Future work will focus on improving scalability using advanced multimodal transformer models and larger datasets. Real-time streaming detection and the integration of explain- able AI techniques will also be explored to enhance trans-parency and system performance.

REFERENCES

1. Y.LeCun,Y.Bengio, and G.Hinton,“ Deep learning,” Nature,vol.521,no. 7553, pp. 436–444, 2015.
2. I.Good fellow et al., “Generative adversarial networks,” in Proc. NIPS,2014.
3. J.Devlin,M.Chang,K.Lee,and K.Toutanova,“BERT:Pre-training of deep bidirectional transformers for language understanding, ”in Proc.NAACL, 2019.
4. H.Nguyen,J.Yamagishi,and I.Echizen,“Capsule-Forensics: Using capsule networks to detect forged images and videos, ”in Proc. ICASSP,2019.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.08

5. Y.Li and S.Lyu, “Exposing deepfake videos by detecting face warping artifacts,” in Proc. CVPR Workshops, 2019.
6. T.Jung, S.Kim, and K.Kim, “Deep Vision: Deepfakes detection using human eye blinking,” IEEE Access, vol. 8, pp. 83144–83154, 2020.
7. M.Afchar, V.Nozick, J.Yamagishi, and I.Echizen, “MesoNet: Acompact facial video forgery detection network,” in Proc. WIFS, 2018.
8. P.Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition,” IEEE Access, vol. 7, pp. 175753–175764, 2019.
9. R.Tolosana et al., “DeepFakes and beyond: A survey of face manipulation and fake detection,” Information Fusion, vol. 64, pp. 131–148,2020.
10. H.Li,B.Li,S.Tan,andJ.Huang,“Identification of deep network generated images using disparities in color components, ”Signal Processing, vol. 174, 2020.
11. S.Wang, O.Wang, A.Owens, R.Zhang, and A.A.Efros, “Detecting photo shopped faces by scripting Photoshop,” in Proc. ICCV, 2019.
12. F.Chollet, “Xception: Deep learning with depth wise separable convolutions,” in Proc. CVPR, 2017.
13. T.Karras,S.Laine,andT.Aila,“A style-based generator architecture for generative adversarial networks,” in Proc. CVPR, 2019.
14. Z.Wuetal.,“ASVspoof:Theautomaticspeakerverificationspoofingchallenge,” in Proc. Interspeech, 2015.
15. H.Tak, J.Patino, M.Todisco, and N. Evans, “End-to-end antispoofing with raw waveform CLDNNs,” in Proc. ICASSP, 2021.
16. A.Agarwaletal.,“Protecting world leaders against deep fakes,”inProc.CVPR Workshops, 2019.
17. A.Dosovitskiyetal.,“An imageisworth16x16words: Transformers for image recognition at scale,” in Proc. ICLR, 2021.
18. A.Radfordetal., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021.
19. A.Vaswanietal., “Attention is allyouneed,”inProc.NIPS,2017.
20. K.Simonyan and A.Zisser man, “Very deep convolutional networks for large scale image recognition,” in Proc. ICLR, 2015.