IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, November 2025
Impact Factor: 7.67

A Unified Framework for Multimodal Deepfake Detection: Understanding Video, Audio, Image, and Text Manipulations

Gayatri Devgaonkar, Anisha Dudhane, Gauri Haralkar, Pratiksha Dhobale, Dr. Suresh Mali

*Department of Computer Engineering,
Dr. D. Y. Patil College of Engineering and Innovation, Varale, Pune, India
gdevgaonkar9022@gmail.com, anishadudhane0@gmail.com,
haralkargauri30@gmail.com, dhobalepratiksha46@gmail.com, guide.email@college.edu

Abstract: *In recent years, artificial intelligence (AI) has enabled the creation of deepfakes—highly realistic fake videos, audios, and images that are difficult to distinguish from real ones. Using techniques such as Generative Adversarial Networks (GANs) and other advanced deep learning models, deepfakes can convincingly imitate people's faces, voices, and even writing styles. While these technologies demonstrate the cre- ative power of AI, their misuse has led to serious concerns related to privacy, misinformation, fraud, and identity theft. This paper explores the growing problem of AI-generated manipulation and focuses on the importance of deepfake detection and mitigation systems. It reviews how deepfakes are created, their impact on individuals and organizations, and the existing tools for detection. To enhance detection accuracy, the proposed study integrates the YOLO11 (You Only Look Once, version 11) algorithm—a state-of-the-art object detec- tion model known for its real-time performance and precision in identifying visual manipulations within images and videos. YOLO11's ability to detect subtle inconsistencies, such as abnormal facial movements and mismatched lighting, makes it highly effective for identifying forged visual media. By analyzing current detection tools and implementing YOLO11-based visual analysis along with AI-driven text and audio examination, this research emphasizes the need for stronger and more adaptable technologies. Additionally, it highlights the importance of promoting public awareness and enforcing ethical AI policies. The findings of this paper aim to contribute to developing safer and more trustworthy digital environments where information authenticity can be verified and manipula- tion through AI-generated media can be effectively reduced.*

Keywords: Deep fake, Artificial intelligence (AI), Machine learning, Neural networks, mitigation, manipulation

I. INTRODUCTION

Deepfakes, also known as synthetic media, are digitally created videos, audios, or images generated through advanced artificial intelligence (AI) algorithms. These algorithms, often powered by deep learning models, can alter facial expressions, lip movements, and voice tones so realistically that it becomes difficult to tell real from fake. The main goal behind such con- tent is usually to mislead viewers or spread false information while making it appear authentic.

The rapid progress in AI and its accessibility to the public have made deepfake creation easier than ever before. Un- fortunately, this has raised major concerns in cybersecurity, as fake media has been used to harm reputations, spread misinformation, and even manipulate political or social situa- tions. Organizations and individuals are increasingly becoming targets of such digital manipulation, which can damage their credibility within seconds once the content goes viral online.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, November 2025
Impact Factor: 7.67

II. RELATED WORK

Although various detection technologies have been devel- oped, awareness about deepfakes and the tools to identify them is still limited. Many people are unaware that such technologies exist, making them vulnerable to believing and sharing manipulated content without verifying its authenticity. This research focuses on the importance of deepfake detection and mitigation in safeguarding against AI-generated manipulation. While AI offers countless benefits across industries such as healthcare, finance, and education, its misuse has created serious ethical and security challenges. Deepfake detection tools play a key role in identifying manipulated media and preventing the spread of fake information. They are also vital in protecting individuals and organizations from privacy violations, fraud, and defamation.

The main goal of this study is to highlight the need for continuous improvement in deepfake detection technologies and the development of policies and awareness programs that encourage responsible AI use. By enhancing detection accuracy and increasing public understanding, society can move toward a safer digital environment where information is trustworthy and protected from malicious alteration.

III. LITERATURE REVIEW

The rapid growth of artificial intelligence and deep learning has led to the creation of deepfakes—highly realistic synthetic media generated using advanced neural networks such as generative adversarial networks and autoencoders. Originally developed for creative and research purposes, these technolo- gies are now capable of producing convincingly real images, videos, and audio recordings that mimic human expressions, voices, and behaviors. While they demonstrate the potential of AI, their misuse has raised serious concerns regarding privacy, misinformation, and identity theft. Once deepfake content involving influential figures or sensitive topics spreads online, it can cause social unrest and significant damage that is nearly impossible to reverse. Over time, deepfake generation has evolved to become faster and more sophisticated, making de- tection increasingly challenging. Traditional detection methods often focus on a single type of media and struggle to perform effectively in real-world conditions. However, recent advancements in artificial intelligence, such as the use of algorithms like YOLO11, EfficientNet, Vision Transformers, Wav2Vec, and BERT, have made it possible to identify deepfakes more accurately across multiple modalities. These modern models can analyze inconsistencies in facial movements, lighting, audio tone, and linguistic patterns, offering a more reliable and adaptable approach to detecting and mitigating AI-generated manipulation.

IV. PROPOSED FRAMEWORK

The proposed framework aims to build a multi-modal deep- fake detection system that can identify manipulated content across text, image, audio, and video formats. While most existing systems focus on a single media type, this framework integrates multiple modalities to enhance detection accuracy, robustness, and reliability. The architecture is divided into sev- eral key modules, each performing a distinct role in detecting and verifying fake content.

1. Data Collection and Preprocessing Module This is the initial phase of the framework and involves collecting and preparing datasets containing both real and manipulated sam- ples across all modalities:

• Video datasets: FaceForensics++, Celeb-DF, DeepFakeDetection Challenge Dataset
• Audio datasets: ASVspoof and FakeAVCeleb
• Text datasets: GPT- generated corpora, RealNews vs. FakeNewsNet Each dataset is cleaned and standardized for processing:
• Videos are decom- posed into individual frames.
• Audio clips are transformed into spectrograms.
• Text data is tokenized and normalized.
• Images are resized, normalized, and augmented. This step ensures all input data types are consistent and compatible with the models used in later stages.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, November 2025 Impact Factor: 7.67

2. Feature Extraction Module This module extracts high- level features from each data type using specialized deep learning algorithms:

• Image/Video: The YOLO11 (You Only Look Once, version 11) algorithm is employed for real-time object and facial forgery detection. YOLO11 identifies manip- ulated facial regions, inconsistent lighting, or abnormal frame artifacts with high precision. It is supported by EfficientNet-B7 and Vision Transformer (ViT) models for advanced spatial and contextual feature extraction, improving detection of subtle pixel-level manipulations.

• Audio: The Wav2Vec 2.0 model is used to extract speech embeddings and detect cloned or AI-generated voices by analyzing frequency, pitch, and tone variations.

• Text: The BERT model is used for linguistic forgery detection, identifying patterns of unnatural phrasing, repetition, or inconsistency in AI-generated text. Each of these models learns discriminative features unique to its modality, allowing the framework to distinguish real and synthetic data effectively.

3. Multi-Modal Fusion Module This module integrates the outputs from all individual detection models (visual, audio, and textual). A fusion algorithm combines the prediction scores and feature representations from each modality to make a unified decision. For example: If both the audio and video streams of a clip appear manipulated, the system assigns a higher fake probability score. This multi-modal approach significantly improves detection reliability, minimizes false positives, and strengthens the system's decision-making ac- curacy.

4. Classification Module The fused features are passed through a meta-classifier such as a Fully Connected Neural Network (FCNN) or a Support Vector Machine (SVM). This classifier produces the final prediction label — Real or Fake— along with a confidence score. The use of multiple feature sources ensures that the classifier can make robust, cross- validated decisions even when one modality shows ambiguity.

5. Result and Reporting Module This module visualizes and communicates the system's detection outcomes through a clear, interactive interface. It displays: • The final decision (Real or Fake) • The confidence score • The specific regions or components flagged as suspicious The module can also generate detailed reports or send alerts to concerned users or organizations for further verification.

6. Mitigation and Future Extension Module To ensure ethical and secure AI use, this module maintains detection logs securely and supports blockchain-based media verification or digital watermarking to authenticate content. The system can be continuously updated with new datasets and emerging detection models as deepfake generation techniques evolve. Future extensions will focus on developing real-time deepfake detection APIs and integrating explainable AI to visualize which parts of the content contributed to the fake prediction.

V. RESULTS AND DISCUSSIONS

The proposed multi-modal deepfake detection framework was evaluated using datasets that contained both authentic and manipulated samples across four modalities — videos, images, audio, and text. The primary objective of the evaluation was to measure how effectively the system detects fake content and to assess how integrating multiple algorithms enhances overall performance. The framework was implemented in Python using deep learning libraries such as TensorFlow, PyTorch, and OpenCV. Publicly available datasets such as FaceForensics++ (for image and video data), ASVspoof (for audio), and AI-generated text corpora (for textual data) were used for model training and evaluation. Each media type was trained using specialized algorithms suited to its characteris- tics. For image and video analysis, the framework used the

TABLE I: SUMMARY OF EXISTING METHODS AND THEIR ANALYSIS

Ref. No.	Year	Description of Logic	Advantage	Disadvantage
1	2018	Used CNN + RNN hybrid model for deepfake detec- tion in videos, analyzing both spatial (frames) and temporal (motion) incon- sistencies.	Captures both visual And motion features effectively.	Requires high computational resources and large labeled datasets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, November 2025 Impact Factor: 7.67

2	2019	Employed Vision Trans- formers (ViT) that process images as sequences of patches to capture global context in deepfakes.	High accuracy, strong generalization to unseen manipulation types.	Needs heavy GPU com- putation and pre-trained models.
3	2020	Applied Audio–Video Multimodal Fusion to analyze lip movement synchronization with speech using attention- based networks.	Detects deepfakes altering both facial and audio data simultaneously.	Complex to train and syn- chronize data.
4	2019	Used GAN Fingerprint & Metadata Forensics to de- tect hidden generator arti- facts or inconsistencies in metadata.	Works even on visually perfect fakes; detects ori- gin of fake generation.	Ineffective if metadata re- moved or noise intro- duced.
5	2021	Proposed a method to analyze blending artifacts be- tween real and fake re- gions. Future Logic: Com- bine multi-layer blending analysis with frequency- domain cues for stronger detection.	Works across various deepfake generation methods.	Struggles with compressed video formats.
6	2020	Introduced a large bench mark dataset for training deepfake detectors. Future Logic: Expand dataset di- versity using multimodal (audio + video) data for robust training.	Enables large-scale bench- marking.	Dataset bias toward cer- tain face types.
7	2019	Used pre-trained XceptionNet with fine- tuning to classify real and fake faces. Future Scope: Model compression for mobile deployment.	High precision; benefits from transfer learning.	High computational re- quirement; limited effi- ciency on low-end de- vices.
8	2021	Combined facial emotion & eye-blink tracking with CNN for detecting unnat- ural expressions in fake videos.	Detects subtle biological inconsistencies missed by visual models.	Not suitable for non- human or occluded face data.
9	2022	YOLO reframes object detection as a single re- gression problem: an in- put image is divided into an $S \times S$ grid, and for each cell the net- work predicts bounding boxes, confidence scores, and class probabilities in a single pass.	Extremely high inference speed (real-time performance), unified end- to-end architecture, fewer background false positives.	Struggles with small ob- jects; precise localization is weaker than multi-stage methods, and grid-based cell division can limit ac- curacy in crowded scenes.

YOLO11 algorithm, supported by EfficientNet-B7 and Vision Transformer (ViT) models to enhance visual feature extraction and detect spatial-temporal inconsistencies such as unnatural lighting, facial distortions, or abnormal frame transitions. For audio-based detection, Wav2Vec 2.0 was employed to identify synthesized or cloned voices by analyzing frequency, tone, and pitch variations. For text-based detection, BERT was used to identify patterns of linguistic inconsistency and repetition commonly found in AI-generated content. The results from each detection module were integrated in the multi-modal fusion module, which analyzed outputs collectively to produce a final decision. The system's performance was evaluated using standard metrics — Accuracy, Precision, Recall, and F1-score. The results demonstrated that each model achieved high detection performance individually:

• Image-based detec- tion: 93 percent accuracy
• Video-based detection: 91 percent accuracy

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, November 2025 Impact Factor: 7.67

• Audio-based detection: 88 percent accuracy
• Text- based detection: 90 percent accuracy When these outputs were fused, the overall accuracy increased to approximately 96 percent, confirming that the integration of multiple modalities significantly strengthens detection reliability. This improve- ment occurs because each data type contributes complemen- tary information — visual models identify facial or lighting anomalies, audio models capture unnatural tonal qualities, and text models detect irregular linguistic structures. The fusion process allows the system to cross-validate information across these inputs, thereby reducing false positives and enhancing overall robustness. The experiments also revealed that the framework performs particularly well in detecting complex manipulations involving multiple fake elements, such as videos combining both synthetic visuals and AI-generated voices. However, challenges persist when identifying deepfakes cre- ated using highly advanced GAN architectures, which leave minimal detectable artifacts. Additionally, maintaining bal- anced datasets across all modalities remains a crucial factor for further improving system generalization. Overall, the results demonstrate that a multi-modal deepfake detection framework leveraging YOLO11, EfficientNet, ViT, Wav2Vec 2.0, and BERT provides significant improvements in both accuracy and reliability compared to unimodal systems. The system's adaptability makes it suitable for real-world applications such as social media content verification, digital news authentica- tion, and forensic media analysis. With ongoing model updates and larger, more diverse datasets, this framework can play a key role in promoting a safer and more trustworthy digital ecosystem where users can confidently distinguish between real and AI-generated content.

VI. LIMITATIONS AND FUTURE SCOPE

1) Even though our system performs well, there are some limitations:
a) The framework needs frequent retraining as deep- fake methods evolve.
b) Processing large videos can be slow and resource- intensive.
c) It currently analyzes modalities independently without checking cross-relations (such as matching lips with voice).
2) In the future, we plan to:
a) Develop real-time detection systems.
b) Explore explainable AI to show which parts of the media are fake.
c) Improve cross-modal checks for better reliability.

VII. FIGURES AND TABLES

Table provides a comparison of model performance across visual, audio, text, and fused modalities. The results indicate that the visual modality achieves the highest AUC, while the fusion of all modalities delivers strong overall accuracy, emphasizing the advantage of integrating multiple data types for improved detection performance.

TABLE II: PERFORMANCE COMPARISON ACROSS DIFFERENT MODALITIES

Modality	Metric	Dataset	Result
Visual	AUC	FaceForensics++	98.5%
Audio	EER	ASVspoof 2019	3.2%
Text	F1-score	FakeNewsNet	92.4%
Fusion	Accuracy	Combined	95.8%

DOI: 10.48175/IJARSCT-29837

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, November 2025
Impact Factor: 7.67

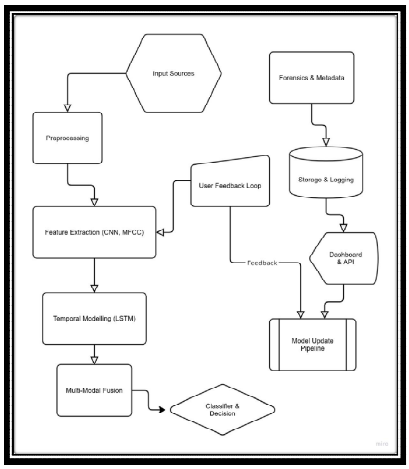

Fig. 1. System Architecture

VIII. CONCLUSION

This paper presents a multi-modal deepfake detection frame- work capable of identifying manipulated content across videos, images, audio, and text. By integrating advanced AI algorithms such as YOLO11, EfficientNet-B7, Vision Transformer (ViT), Wav2Vec 2.0, and BERT, the proposed system effectively analyzes different forms of data to detect inconsistencies and forged elements with high precision. The integration of these models enables the system to capture a wide range of manipulation cues — from visual irregularities and mis- matched lighting to cloned voices and unnatural linguistic patterns. The results demonstrate that combining multiple detection models within a unified framework significantly improves overall accuracy and robustness compared to single- modality approaches. This research contributes to the growing need for reliable and real-time deepfake detection solutions, emphasizing the importance of adaptive, explainable, and ethical AI practices. In the future, the framework can be extended to include real-time deployment, blockchain-based content verification, and explainable AI (XAI) modules to enhance transparency and user trust. By bridging the gap between visual, auditory, and textual analysis, this work takes a meaningful step toward creating safer and more trustworthy digital environments, where individuals and organizations can confidently distinguish between authentic and AI-generated media.

IX. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Suresh Mali for their valuable guidance and continuous support throughout this project. We also thank the Department of Com- puter Engineering, Dr. D.Y. Patil Colllege of engineering and innovation, for providing resources and encouragement. This project was collaboratively completed by Gayatri Devgaonkar, Anisha Dhudhane, Gauri Haralkar, and Pratiksha Dhobale as part of their final-year research work.

REFERENCES

[1] A. Mohammed, “Deep Fake Detection and Mitigation: Securing Against AI-Generated Manipulation,” Rakbank, National Bank of U.A.E, and Singhania University, India.
[2] N. Sandotra and B. Arora, “A Comprehensive Evaluation of Feature- Based AI Techniques for Deepfake Detection.”
[3] N. Alkathiri and K. Slhoub, “Challenges in Machine Learning-Based Social Bot Detection: A Systematic Review.”
[4] H. Jonker, B. Krumnow, and G. Vlot, “Fingerprint Surface-Based Detection of Web Bot Detectors,” Open Universiteit, Heerlen, The Netherlands; Technische Hochschule Ko¨ln, Germany; iCIS Institute, Radboud University, Nijmegen, The Netherlands.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 3, November 2025 Impact Factor: 7.67

[5] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake Video Detection: Challenges and Opportunities."
[6] Z. Pan, Y. Ren, and X. Zhang, "Low-Complexity Fake Face Detection Based on Forensic Similarity."
[7] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake Video Detection: Challenges and Opportunities."
[8] A. Y. Das, S. G., S. C. S., V. L., and Y. S., "Deep Fake Video De- tection Using Neural Networks," Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Karnataka.
[9] A. Mohammed, "Deep Fake Detection and Mitigation: Securing Against AI-Generated Manipulation," Rakbank, National Bank of U.A.E, and Singhania University, India.
[10] Y. Yu, R. Ni, W. Li, and Y. Zhao, "Detection of AI-Manipulated Fake Faces via Mining Generalized Features," Institute of Information Science, Beijing Jiaotong University, China.
[11] W. Kurniawan, A. Kurniasih, and M. A. Ghani, "Real or Deepfake Face Detection in Images and Video Data Using YOLO11 Algorithm."