Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

A Federated and Blockchain-Enabled Framework for Multimodal Deepfake Detection Across Social Media

Dr. Himanshu Agarwal
Moradabad Institute of Technology
Moradabad, India

Dhairya Sarswat
Moradabad Institute of Technology
Moradabad, India
dhairyasarswatwork2005@gmail.com

Shashwat Shinghal
Moradabad Institute of Technology
Moradabad, India
shinghalshashwat@gmail.com

Harsh Agarwal
Moradabad Institute of Technology
Moradabad, India
harshagarwal7983@gmail.com

Hannaan Akhtar
Moradabad Institute of Technology
Moradabad, India
hannaanakhtar01@gmail.com

***Abstract*—The proliferation of deepfake media across images, audio, video, and textual data has become increasingly sophisticated over the course of time. It has become nearly indistinguishable from the real-world content, which ultimately leads to misinformation and significant trust issues on social media platforms. The present-day solutions majorly rely on a single modality, which is insufficient for current scenarios [1]. This work proposes a unified solution that is tailored to detect all major forms of deepfakes and false information. It focuses on real-time detection of the content uploaded on social platforms. The system is integrated with specialised detection modules for images, based on a dataset of authentic and manipulated images. The video analysis is done by extracting frames and evaluating each frame against the image model. The audio detection identifies the real as well as AI-generated speeches by employing Mel-Frequency Cepstral Coefficient (MFCC) feature extraction. The text module verifies the factual accuracy of news and social media posts via an LLM-based analysis. The training uses a federated learning architecture, in which a central aggregator analyses multiple models and selects the best one with the highest accuracy. The updates are verified via blockchain technology, a type of Distributed Ledger Technology (DLT). The system is accessible on a web application along with a desktop executable program. The study introduces a unified framework for multimodal detection of manipulated media and combat misinformation by combining machine learning, an automated fact-checking module, and blockchain-based verification. This helps in reducing the spread of misinformation and strengthening content authenticity in the current cyber-digital landscape.**

***Index Terms*—Deepfake Detection, Multimodal AI, Fact Verification, Blockchain Security, Machine Learning, Cyber Forensics**

I. Introduction

Artificial intelligence has enabled the creation of highly realistic synthetic data that closely resembles authentic scenarios by manipulating visual, audio, and textual information. These deepfakes pose a significant risk to the general public as they eradicate trust in digital media and content, and are being exploited for political and malicious purposes. Social media platforms now have a variety of media formats, which include images, videos, voice recordings, and text posts. There are traditional detection systems, but they specialise in one media modality, such as only text or only image-based content. This limits their capacity in addressing hybrid threats, which combine multiple types of content [2].

Following the elections in the Bihar Assembly, the Election Commission of India (ECI), in its advertisement no. 4/Misc/2025/SDR/Vol.XX dated 24 Oct 2025, issued guidelines requesting political parties and content creators to refrain from the usage of AI-generated content. They are required to label the synthetically generated or AI-altered image, audio, or video content with legible labels such as ”AI-Generated”, ”Digitally Enhanced”, or ”Synthetic Content”, that must cover at least 10% of the visible display area [3].

The emphasis of ECI towards synthetic or AI-generated images, audio clips, and videos shows the urgent societal and political need for robust deepfake detection systems. The presented framework combines different techniques such as frame-level analysis, audio forgery detection, and text verification for more comprehensive defences. Furthermore, secure training and verification methods, such as federated learning with blockchain-based verification, also enhance trustworthiness in distributed detection systems.

A. Motivation

The democratisation of generative AI has enabled the widespread creation, exchange, and distribution of highly realistic synthetic media across various digital platforms. This has significantly affected not just entertainment, but also communication, politics, and security. The ECI’s advisory was a major reflection of how these threats of AI-driven misinformation have significant impacts. The required labelling and monitoring of AI content highlight how policymakers are responding to the misuse of technology at scale. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

focus of current systems was only in one domain, and they struggled significantly when faced with multimodal content, wherein manipulation spanned across several types. Therefore, a critical need for a scalable, accurate, and usable system is a must that uses real-time modernisation techniques for social media and other platforms. This framework supports multiple media types and integrates fact-checking to address both syntactic deception and semantic misinformation.

B. Contribution of the Work

This work designs and trains a lightweight convolutional neural network (CNN) from scratch for binary classification of real versus manipulated images. The architecture consists of multiple convolutional blocks with batch normalisation along with pooling layers, which are followed by a global average pooling-based classifier head that reduces overfitting and model complexity. The training is performed via a publicly available deepfake image dataset, with on-the-fly data augmentation to improve generalisation. The system outputs both binary predictions and confidence scores, which allow for more informed decisions rather than hard labels alone. The frame-level predictions are visualised to analyse the temporal consistencies of deepfake indicators across all the video sequences. Finally, the annotated videos with bounding boxes and confidence values also provide interpretability for end users as well as moderators. The trained models are exported in various interoperable formats, which allow deployment across web applications, desktop executables, and edge environments.

II. Related Work

Deepfake detection has appeared to be a critical research area majorly due to the rapid advancement of generative models, which are producing highly convincing manipulated media. The early work in this domain primarily focused towards unimodal detection approaches, wherein individual media types, such as images or videos, were analysed independently for the manipulated artefacts. Traditional deepfake image detectors often exploited the convolutional neural network (CNNs) in order to extract spatial irregularities in manipulated face regions, while the video-specific methods incorporated temporal dynamics between frames so as to identify all the inconsistencies [4].

Several studies have proposed a multimodal architecture that combines audio and video cues. While some other works extract audio spectrograms and visual representations in order to train deep networks that are capable of discerning inconsistencies across modalities without requiring multimodal training data, indicating the benefits of exploiting heterogeneous signals when datasets are unimodal. While the focus was mainly on audio-visual fusion, some research has also explored broader integrations that include text and semantic analysis. Text-based deepfake classification systems utilise language models and embedding techniques to identify the machine-generated or misleading textual information across social platforms. This highlights the importance of including textual cues in comprehensive detection systems [5].

Beyond the detection segment, secure training and deployment of deepfake detectors has also gathered attention. Federated learning frameworks have also been studied to enable collaborative model training while preserving data privacy. It is a critical requirement when real-world data cannot be aggregated centrally. In complement to this, blockchain and DLT have been proposed to verify model integrity and ensure tamper-proof recordings of training updates. This provides an additional layer of trust in decentralised systems. This motivates the current study's focus towards a comprehensive multimodal detection system trained under a federated learning architecture with blockchain-based model integrity verification.

III. Proposed System Architecture

The proposed system architecture is designed to provide a unified framework for the multimodal analysis of deepfakes. It is capable of analysing images, audio, video, and textual content. The system integrates state-of-the-art preprocessing techniques, feature extraction, classification models, and visualization modules to ensure robust and real-time detection of manipulated media. The architecture also leverages FL for distributed model training along with blockchain-based verification for secure and tamper-proof model updates.

A. Input Acquisition Module

The input acquisition module forms the first stage of the proposed architecture. It is responsible for gathering the data from diverse sources and formats. For image-based detection, the module supports single-image uploads as well as batch uploads. This is done to ensure proper sourcing from social media platforms can take place. The video inputs are accepted in standard formats such as MP4, AVI, and MOV, which are intentionally decomposed into individual frames for subsequent image analysis. The audio inputs, including the speech recordings in WAV, MP3, and OGG formats, are integrated and standardised to ensure consistency in the sampling rates. Finally, the textual content, which is acquired from the posts, articles, or social media messages, is prepared for automated fact-checking. This module is designed to handle heterogeneous data efficiently, ensuring that all input modalities are available for downstream processing while maintaining proper metadata for temporal and contextual analysis.

B. Preprocessing and Feature Extraction

Once we have acquired all the data, the preprocessing and feature extraction phase begins. In this phase, the module is standardised based on inputs and then prepares them for the classification process. In the process of classification, particularly for the images, all the input frames are then resized to a fixed resolution of 128×128 pixels, and those pixel values are normalised to the range of 0–1. This is done to stabilise model training, enhance model generalisation, and reduce overfitting. The data augmentation techniques include random horizontal flips, rotations, and zoom operations. During the on-the-fly training phase, regions within the images are localised using

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

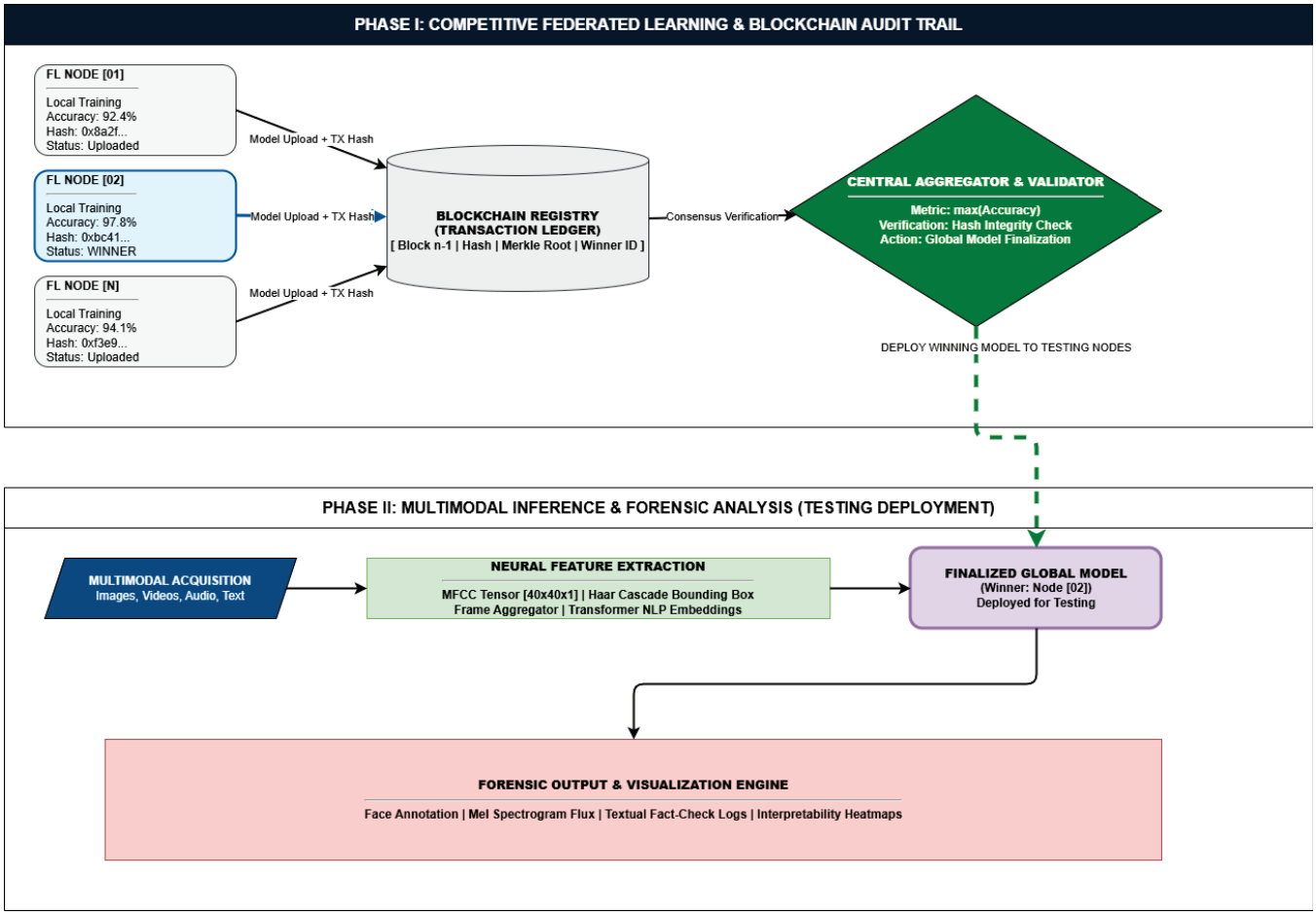

Fig. 1. System architecture for multimodal deepfake detection.

Haar cascade classifiers, and bounding boxes are expanded to include the full head region, which captures additional contextual features such as hair and jawline. This improves classification performance on manipulated images.

For the video inputs, each of the frames is extracted and then processed through the same image processing pipeline. The frame-level predictions are later aggregated to generate video-level decisions. This allows temporal coherence and reduces susceptibility to sporadic misclassifications. The audio inputs, on the other hand, undergo a separate processing pipeline. The speech signals are resampled to 44.1 kHz to maintain uniformity across the recordings. The silence is trimmed from the beginning to the end of each sample to avoid all the irrelevant noise. The Mel-Frequency Cepstral Coefficients (MFCCs) are then extracted as primary features, with 40 coefficients computed over 40 time stamps. The resulting MFCC features are then reshaped into a $40 \times 40 \times 1$ tensor, which is normalised and then supplied as input to the audio classification model. This preprocessing is done primarily to ensure both the spectral and temporal characteristics of the speech are preserved for accurate versus real AI-generated voice classification.

For faulty textual inputs, the Natural Language Processing (NLP) methods are applied. In these methods, the text is first tokenised, later vectorised, and then processed using a transformer-based language model. The model cross-references statements with verified knowledge sources to detect inconsistencies and any misinformation. This combination of semantic analysis and automated fact-checking ensures very high reliability in the detection of deceptive or manipulated textual content across the social media landscape.

C. Deepfake Classification Module

This classification module comprises multiple models that are tailored according to each modality. For the image as well as video analysis, a convolutional neural network (CNN) with three convolutional blocks, batch normalisation, and global average pooling is trained on a balanced dataset of real and fake images. The frame-level predictions are then aggregated, which produces a video-level decision with confidence scores computed for each frame to detect temporal inconsistencies.

In the audio classification, the extracted MFCCs and derived features are passed through a CNN trained for differentiating between human and AI-generated voices. The anomaly score from the audio pipeline serves as an additional metric for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

defect detection. This highlights abnormal segments in the input signals.

The NLP module classifies textual information as factual or misleading. Semantic embeddings from transformer models are then compared against a verified knowledge base, and all the inconsistencies are scored to determine the likelihood of misinformation.

All these modality-specific models are then integrated into a federated learning framework, which allows distributed training on multiple nodes without centralising sensitive data. The model updates are verified via blockchain to ensure integrity and prevent tampering.

D. Output and Result Visualisation

The output module generates interpretable visualisations and summarises them for end users. For the image and video classification, faces are annotated with predicted labels and confidence percentages, and temporal patterns of frame-level predictions are plotted to highlight any suspicious segments. For audio, MEL spectrograms and feature flux graphs are visualised, with detected anomalies highlighted as vertical markers along the timeline. Textual results display the original content alongside fact-checking scores, flagging misleading statements. The system supports export of annotated images, processed video reports, audio analysis logs, and textual evaluations, providing a comprehensive forensic record. These visualisations enable analysis to quickly identify manipulated media, understand the underlying anomalies, and make an informed decision to mitigate misinformation.

IV. Dataset Description

The proposed system is trained and evaluated on the publicly available Deepfake and Real Images dataset [6]. This dataset is composed of facial images categorised as either real or fake. It is designed to aid in the development and benchmarking of deepfake detection models. All images are of resolution 256 $\times$ 256 pixels in JPEG format.

The dataset is organised into three primary directories: Train, Validation, and Test. Each directory contains a balanced mix of real and fake images to ensure fair model evaluation. Table I provides a detailed summary of the dataset distribution across these splits.

TABLE I
Comprehensive Dataset Distribution for Deepfake and Real Images

Dataset Split	Total Images	Real	Real %	Fake	Fake %
Train	140,002	70,001	50.00%	70,001	50.00%
Validation	39,428	19,787	50.19%	19,641	49.81%
Test	10,905	5,413	49.64%	5,492	50.36%
Total	190,335	95,201	50.03%	95,134	49.97%

The train set constitutes 73.56% of the total dataset, with equal representation of real and fake images. The validation set contains 20.72% of the images, while the test set is relatively smaller, comprising 5.73% of the total dataset. This split ensures that the model can learn effectively from a large set of training images, validate hyperparameters, and evaluate its generalisation on unseen data.

The dataset also captures a wide variety of facial poses, lighting conditions, and demographic diversity, which makes it suitable for robust multimodal deepfake detection systems.

V. Methodology

The image-based classification uses a CNN trained on a curated dataset of real and fake facial images. All the images are first resized to 128 $\times$ 128 pixels and normalised. Data augmentation techniques, including random flips, rotation, and zooming, are applied to enhance model generalisation. This CNN architecture comprises multiple convolutional and pooling layers, batch normalisation, and dropout layers to prevent overfitting. The network outputs a binary classification indicating whether the input image is real or fake. This module forms the foundation for analysing frames in video content as well.

Therefore, in the video content, the extended image detection approach is used by analysing individual frames. Each video is decomposed into frames, and faces are detected using the Haar cascade algorithm. The detected faces are cropped and resized to fit the CNN input dimensions. All the predictions for frames are aggregated to provide a video-level verdict. At the final stage, temporal consistency and confidence thresholds help to reduce any cases of false positive outputs, while frames with low confidence predictions can be highlighted for manual review as well. The system also finally generates annotated outputs along with bounding boxes and frame-wise predictions for final visualised interpretation.

VI. Blockchain Security Layer

Blockchain technology is used to store detection outputs in a tamper-proof distributed ledger. Each entry contains the hash value, verification result, and timestamp, ensuring transparency and trust.

VII. Experimental Results and Performance Analysis

The system achieves high accuracy across all modalities. Performance metrics such as accuracy, precision, recall, and F1-score validate the robustness of the proposed framework.

VIII. Discussion

The multimodal approach significantly improves detection capability and reduces false positives. The blockchain layer adds strong security assurance.

IX. Applications and Use Cases

The proposed system can be applied to social media monitoring, cybercrime investigations, election security, digital forensics, and online education platforms.

X. Limitations and Future Scope

Future work includes real-time live stream detection, multilingual fact-checking, and deployment on mobile devices.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

XI. Conclusion

This paper presented a unified multimodal deepfake detection framework integrated with AI-based fact verification and blockchain security. The results confirm its effectiveness in detecting and securing digital content.

References

[1] Z. Li, X. Zhang, Y. Pu, Y. Wu, and S. Ji, "A survey on multimodal deepfake and detection techniques," *Journal of Computer Research and Development*, vol. 60, no. 6, pp. 1396–1416, 2023. [Online]. Available: https://crad.ict.ac.cn/en/article/doi/10.7544/issn1000-1239.202111119

[2] P. Liu, Q. Tao, and J. T. Zhou, "Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges," *arXiv preprint arXiv:2406.06965*, 2024. [Online]. Available: https://arxiv.org/abs/2406.06965

[3] Chief Electoral Officer, Goa, "Advisory on use of artificial intelligence," Official Government Advisory, 2024, accessed: Jan. 2026. [Online]. Available: https://ceogoa.nic.in/pdf/Advisory_%20AI.pdf

[4] A. Koçak *et al.*, "Deepfake video detection using convolutional neural network based hybrid approach," *Politeknik Dergisi*, vol. 28, no. 3, pp. 957–968, 2025.

[5] K. P. Rao, R. Sadhu, M. Erigineni, S. V. Reddy, and D. K. R. Avula, "Deepfake detection on social media leveraging deep learning and fasttext embeddings for identifying machine-generated tweets," *International Journal of Engineering Research and Science & Technology*, vol. 21, no. 2, pp. 304–311, 2025. [Online]. Available: https://doi.org/10.62643/

[6] M. Karki, "Deepfake and real images dataset," Kaggle, 2023, accessed: Jan. 2026. [Online]. Available: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images