IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEVELOPMENT OF A UNIFIED MULTIMODAL SYSTEM FOR DETECTING MANIPULATED AUDIO, IMAGE, AND VIDEO CONTENT

[1]Balakrishna Tilakachuri, [2]Mohammed Adil, [3]Motukuri Shrivalli, [4]Matta Kanakasri, [5]Parasa Naga Veera Vardhan

[1]Assistant Professor, [2,3,4,5] Student

Department of Computer Science and Engineering,

Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India

Abstract: The ability to generate convincing synthetic audio, image, and video content has introduced significant challenges for the verification of digital media authenticity. Deepfake manipulation techniques can closely replicate real-world media characteristics, rendering single-modality and manual detection approaches ineffective in many practical scenarios. To address this issue, this work implements a multimodal deepfake detection system that processes audio, image, and video inputs through independent yet coordinated analysis pipelines. The proposed framework adopts modality-specific preprocessing and feature extraction strategies, followed by dedicated classification models tailored to the characteristics of each media type. Audio inputs are analysed using a machine learning–based classifier, while image and video content are evaluated using convolutional neural networks designed to capture spatial and frame-level manipulation artifacts. Experimental evaluation on benchmark datasets confirms the system's ability to accurately differentiate authentic content from manipulated media across all modalities. The modular structure of the framework supports extensibility and makes it suitable for deployment in forensic analysis, media verification, and security-oriented applications.

Index Terms - Deepfake Detection, Audio Deepfake, Image Deepfake, Video Deepfake, Deep Learning, Machine Learning, Multimedia Forensics.

I. INTRODUCTION

Recent progress in generative modelling has enabled the automated creation of manipulated digital media that can closely imitate authentic audio, image, and video content. These manipulated samples, commonly referred to as deepfakes, pose practical challenges for digital content verification due to their visual and auditory realism. As manipulated media becomes easier to produce and distribute, the need for reliable detection mechanisms has become increasingly important in multimedia security and forensic analysis [1], [2]. Earlier media verification techniques were primarily designed to identify conventional editing artifacts or basic tampering operations. Such methods are often ineffective against manipulation techniques driven by deep neural networks, which introduce fewer detectable inconsistencies [3]. In response, learning-based detection methods have been proposed to identify subtle artifacts introduced during media synthesis, leveraging supervised learning techniques for improved classification performance [4], [31]. However, many of these approaches focus on a single modality, limiting their applicability in scenarios where manipulated audio, image, and video content coexist [4]. Addressing these limitations requires a detection framework capable of handling multiple

media modalities while preserving the distinct characteristics of each input type. This work is motivated by the need for a practical multimodal detection system that balances detection accuracy, architectural simplicity, and deployment feasibility. The proposed approach focuses on independent modality-specific analysis combined within a unified system architecture, enabling robust detection without excessive computational complexity. Early deepfake detection research primarily focused on identifying visual artifacts in manipulated images and videos. Image-based approaches typically analyse spatial inconsistencies such as unnatural textures, distorted facial landmarks, or irregular lighting patterns using convolutional neural networks [4], [5]. Video-based detection methods extend this analysis by incorporating temporal information across frames to capture motion inconsistencies introduced during manipulation [6]. In parallel, the rise of neural text-to-speech and voice cloning technologies has made audio deepfake detection an equally important problem. Synthetic speech often contains subtle spectral and temporal irregularities that can be identified using signal processing techniques and machine learning classifiers [11]. While several effective detection methods have been proposed for individual modalities, most existing systems are limited to either audio, image, or video analysis. In real-world scenarios, deepfake content may involve multiple manipulated media types simultaneously, such as a fake video accompanied by synthetic audio. This limitation motivates the need for a unified detection framework capable of handling multiple modalities while preserving the unique characteristics of each. This paper addresses this gap by proposing a multimodal deepfake detection system that independently processes audio, image, and video inputs using specialized models within a single integrated architecture.

II. LITERATURE REVIEW

Deepfake detection research has evolved alongside advances in multimedia forensics and generative modelling techniques. Early forensic approaches relied on handcrafted features and statistical analysis to detect visual inconsistencies introduced during image manipulation. While effective for traditional tampering, these methods were found to be insufficient against deep learning–based synthesis techniques [5]. Image-based deepfake detection methods later adopted convolutional neural networks to identify artifacts introduced during face swapping and facial synthesis. Benchmark datasets such as FaceForensics++ enabled systematic evaluation of these approaches under controlled conditions [6]. Despite strong performance on curated datasets, many image-based detectors exhibited reduced generalization when evaluated on compressed or low-quality media [7]. Video deepfake detection techniques extended image-based analysis by incorporating temporal information across frames. Recurrent neural networks and frame aggregation strategies were introduced to capture motion-related inconsistencies caused by manipulation [8]. Although these approaches improved robustness, they often required substantial computational resources and large annotated datasets [9]. In parallel, audio deepfake detection research focused on identifying artifacts introduced during voice cloning and text-to-speech synthesis. Spectral representations and cepstral features have been widely used to characterize subtle irregularities in synthesized speech signals [10], [11]. However, audio-only detection systems remain vulnerable when manipulated media combines multiple modalities. To address these challenges, recent studies have explored multimodal detection strategies that integrate information from audio, image, and video sources [12]. While such approaches improve detection coverage, they frequently introduce increased architectural complexity. These observations motivate the development of a modular multimodal framework that balances detection performance, interpretability, and practical deployment requirements. Multimodal deepfake detection systems attempt to combine audio and visual cues to improve robustness. While such systems demonstrate improved performance, they often introduce architectural complexity and require synchronized multimodal data [12], [15]. Based on the limitations identified in existing approaches, a modular multimodal system that independently processes each modality offers a practical balance between performance and complexity. Supervised learning approaches have also been successfully applied in various classification tasks, demonstrating strong predictive capabilities across different domains [31].

III. METHODOLOGY

The proposed deepfake detection methodology is designed around independent processing pipelines for audio, image, and video modalities. This modular design allows each media type to be analysed using techniques best suited to its characteristics while maintaining a unified system workflow.

A. Audio Deepfake Detection

Audio inputs undergo preprocessing steps including resampling, normalization, and segmentation. From each audio segment, spectral and temporal features such as MFCCs, spectral centroid, and energy-related descriptors are extracted. These features are then classified using a Random Forest algorithm, which is chosen for its robustness to feature variation and resistance to overfitting [16], [30].

B.Image Deepfake Detection

Image inputs are first processed using face detection to isolate facial regions. The extracted faces are resized and normalized before being passed to a convolutional neural network. The CNN automatically learns discriminative spatial features that highlight manipulation artifacts present in deepfake images [19].

C. Video Deepfake Detection

Video inputs are decomposed into frames at fixed intervals. Facial regions are detected in each frame and analysed using a CNN-based classifier. Frame-level predictions are aggregated using a majority voting strategy to determine the final video classification. This approach captures both spatial and temporal inconsistencies without requiring complex temporal models [6], [13].

IV. SYSTEM ARCHITECTURE

The proposed system adopts a modular and layered architecture designed to efficiently handle multiple media modalities. The overall framework is composed of four primary layers: input, preprocessing, detection, and output. Initially, media files provided by the user are received through the input layer, where basic validation checks are performed to ensure format compatibility and data integrity.Following validation, the data is directed to modality-specific preprocessing modules. These modules perform essential transformations such as resizing, normalization, and feature preparation to ensure consistency with the training conditions of the detection models. Each media type, including image, video, and audio, is processed independently to preserve modality-specific characteristics.In the detection layer, preprocessed inputs are passed to dedicated deep learning models trained for each modality. These models perform binary classification to determine whether the input is real or manipulated, while also generating a confidence score to indicate prediction certainty.Finally, the output layer aggregates the results and presents them in a unified and interpretable format. The modular design of the system ensures flexibility, allowing individual detection components to be updated or replaced without affecting the overall workflow. This architecture supports scalability and facilitates future extensions toward fully integrated multimodal deepfake detection systems.

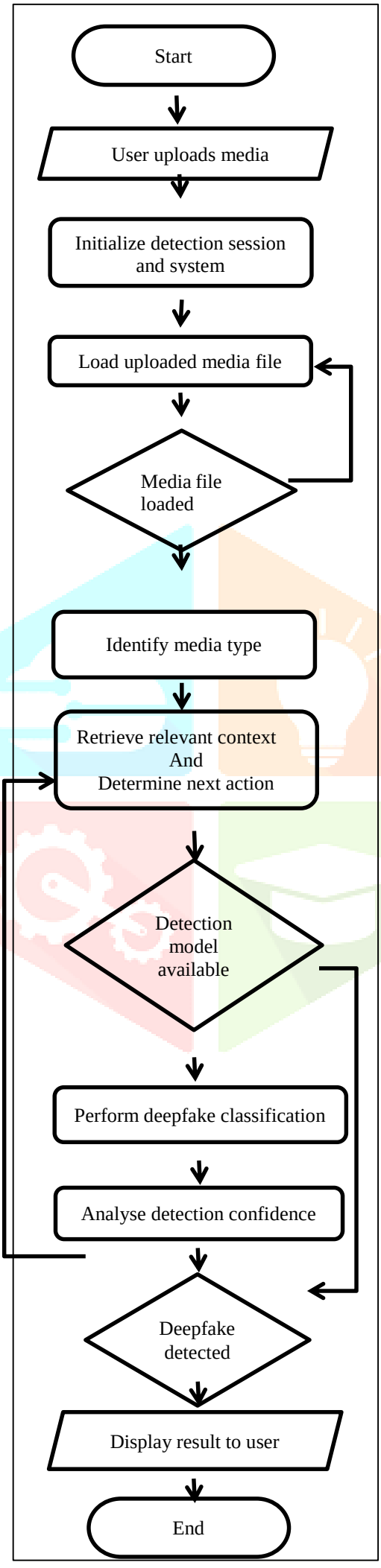
Start
User uploads media
Initialize detection session and system
Load uploaded media file
Media file loaded
Identify media type
Retrieve relevant context And Determine next action
Detection model available
Perform deepfake classification
Analyse detection confidence
Deepfake detected
Display result to user
End

Fig 1: Workflow of Context-Driven Web Task Automation

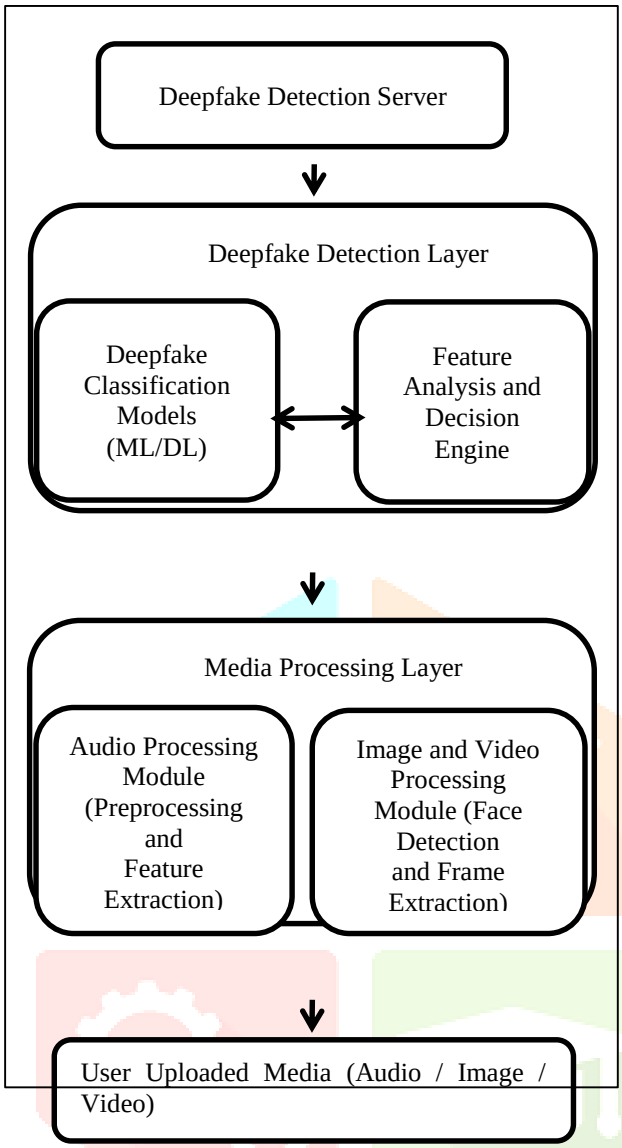

Fig 2: System architecture of the proposed deepfake detection system

V. RESULTS AND DISCUSSION

The proposed system was evaluated using benchmark datasets for audio, image, and video deepfake detection. Experimental results indicate that domain-specific detection pipelines improve classification accuracy and interpretability compared to unified models that attempt to handle all media types simultaneously. Audio deepfake detection demonstrated effective discrimination between genuine and synthetic speech, particularly for voice cloning attacks. Image detection successfully identified facial manipulation artifacts, while video detection benefited from temporal analysis, enabling improved detection of motion inconsistencies. Despite promising results, performance varied depending on dataset quality, compression level, and manipulation technique. These observations align with findings reported in previous studies, which highlight the challenges of generalization in deepfake detection [6], [10].

A. Overall System Output Interface

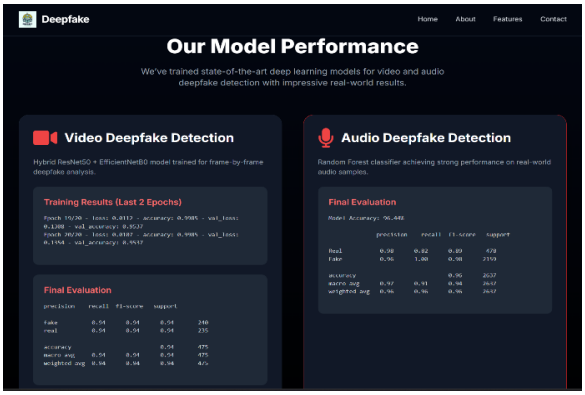

Fig 3: Web-based interface showing the overall performance of the proposed deepfake detection system.

B. Audio Deepfake Detection – Classification Report

Class	Precision	Recall	F1-score	Support
Real	0.98	0.82	0.89	478
Fake	0.96	1.00	0.98	2159
Accuracy			0.96	2637
Macro Avg	0.97	0.91	0.94	2637
Weighted Avg	0.96	0.96	0.96	2637

Table I: Performance evaluation of the audio deepfake detection model using precision, recall, F1-score, and accuracy.

C. Confusion Matrix for Audio Deepfake Detection

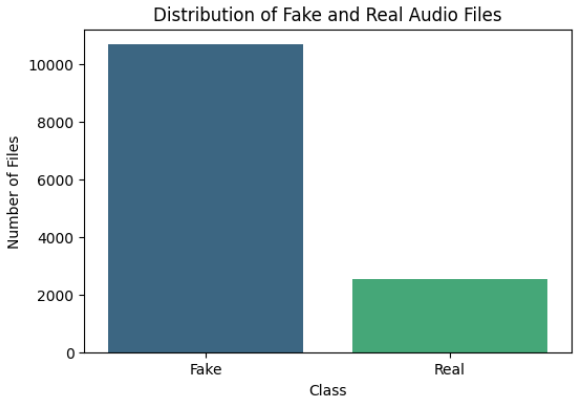

Fig 4: Confusion matrix for the proposed audio deepfake detection model.

D. Video Deepfake Detection – Frame-by-Frame Analysis

Frame-level analysis of video inputs shows that the system successfully localizes facial regions and assigns appropriate classification labels. Genuine videos are consistently classified as real, while manipulated videos exhibit frame-level predictions indicating deepfake artifacts. This qualitative analysis enhances interpretability and supports forensic investigation.

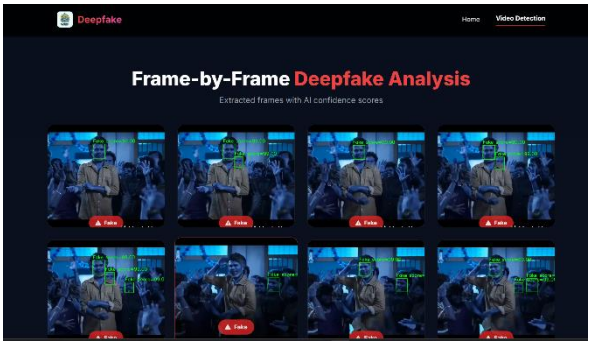

Fig 5: Frame-by-frame deepfake detection results showing facial localization and confidence scores for manipulated video content.

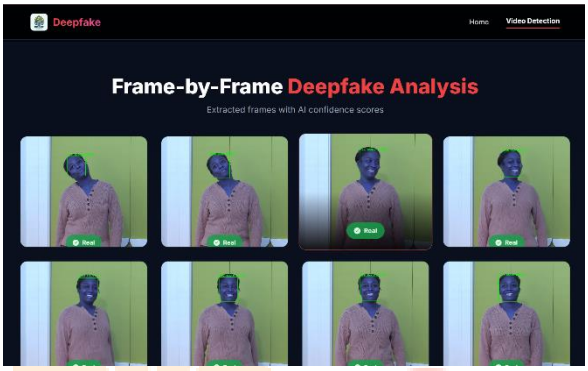

Fig 6: Frame-level analysis of genuine video samples correctly classified as real by the proposed system.

E. Video Deepfake Detection – Classification Report

Classification reports provide a detailed evaluation of model performance by presenting precision, recall, and F1-score for real and fake classes. High precision values indicate a low false positive rate, while strong recall values demonstrate effective detection of manipulated samples. Balanced F1-scores across classes confirm the reliability of the proposed system.

Class	Precision	Recall	F1-score	Support
Real	0.94	0.94	0.94	240
Fake	0.94	0.94	0.94	235
Accuracy			0.94	475
Macro Avg	0.94	0.94	0.94	475
Weighted Avg	0.94	0.94	0.94	475

Table II: Classification report of the video deepfake detection model showing precision, recall, F1-score, and support for real and fake classes.

F. Accuracy

The accuracy results demonstrate that the proposed system effectively learns discriminative patterns for deepfake detection across all modalities. The steady improvement in accuracy during training indicates stable learning behavior. High accuracy values achieved on test data confirm that the selected models generalize well to unseen samples and are capable of distinguishing real media from manipulated content.

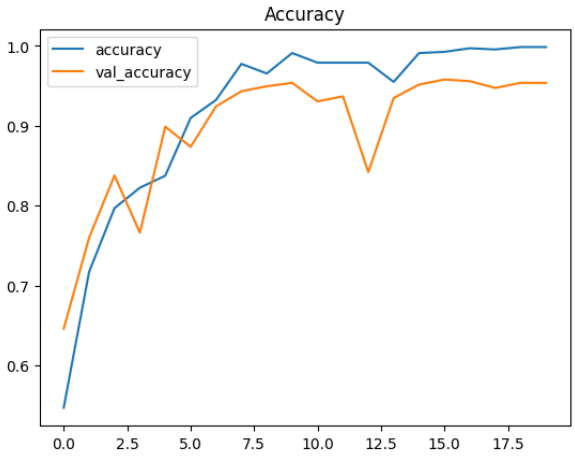

Fig 7: Training and validation accuracy of the video deepfake detection model across multiple epochs.

G. Loss

The loss curves illustrate the convergence behavior of the detection models. A consistent decrease in loss values across training epochs indicates effective optimization and parameter learning. The minimal gap between training and validation loss suggests that the models maintain a good balance between bias and variance, reducing the risk of overfitting.

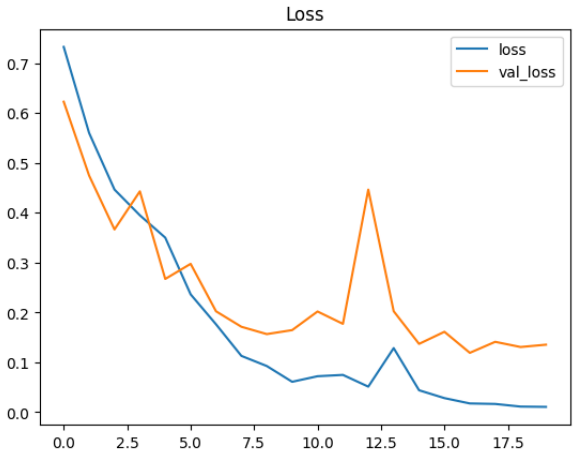

Fig 8: Training and validation loss of the video deepfake detection model during the learning process.

Algorithm: Random Forest–Based Audio Deepfake Detection

Input:
Audio dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
where $x_i$represents extracted audio features and

$$y_i \in \{Real, Fake\}$$

Output:
Predicted class label $\hat{y} \in \{Real, Fake\}$
Step 1: Audio Preprocessing
1. Convert all audio samples to a uniform sampling rate.
2. Remove noise and normalize amplitude.
3. Segment audio signals into fixed-duration frames.

Step 2: Feature Extraction
From each audio sample, extract:

- Mel-Frequency Cepstral Coefficients (MFCCs)
- Spectral centroid
- Spectral bandwidth
- Zero-crossing rate
- Root Mean Square (RMS) energy

Each audio sample is represented as a feature vector:

$$x_i = [f_1, f_2, \dots, f_m]$$

Step 3: Bootstrap Sampling

1. Generate multiple bootstrap samples from the training dataset.
2. Each bootstrap sample is created by random sampling with replacement.

Step 4: Decision Tree Construction

For each bootstrap sample:

1. Construct a decision tree.
2. At each node:
   - Randomly select a subset of features.
   - Choose the best split using impurity measures.
3. Grow the tree until a stopping condition is met.

Step 5: Forest Formation

Repeat Step 4 to build T independent decision trees, forming a Random Forest.

Step 6: Classification

For a new audio sample:

1. Pass the feature vector through all decision trees.
2. Each tree outputs a class prediction.
3. Final prediction is obtained using majority voting.

Step 7: Output

Return the predicted label:

$$\hat{y} = \text{mode}\{y_1, y_2, \dots, y_T\}$$

MATHEMATICAL FORMULATION / FORMULAE

1. Feature Vector Representation

Let an audio sample be represented as:

$$x = [MFCC_1, MFCC_2, \dots, MFCC_k, ZCR, RMS, SC]$$

where:

- $MFCC$= Mel-Frequency Cepstral Coefficients
- ZCR= Zero-Crossing Rate
- RMS= Root Mean Square Energy
- SC= Spectral Centroid

2. Gini Impurity (Used for Splitting)

Random Forest commonly uses Gini impurity to evaluate splits:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

where:

- C= number of classes
- p_i= probability of class i

The split that minimizes Gini impurity is selected.

3. Information Gain (Alternative Criterion)

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

where:

- $H(S)$= entropy of dataset
- A= selected feature
- S_v= subset after split

Entropy:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

4. Majority Voting (Final Prediction)

Given predictions from T trees:

$$\hat{y} = \arg \max_{c \in \{Real, Fake\}} \sum_{t=1}^{T} I(y_t = c)$$

where:

- $I(\cdot)$is the indicator function
- $y_t$is the prediction of the t^{th}tree

5. Classification Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

6. Precision, Recall, F1-score

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

VI. CONCLUSION

This study presents the design and implementation of a multimodal deepfake detection framework capable of analyzing image, video, and audio inputs within a unified system. The approach leverages modality-specific preprocessing techniques and dedicated deep learning models to capture manipulation artifacts unique to each data type. Experimental results from the image-based detection model demonstrate reliable performance in distinguishing real and manipulated content, validating the effectiveness of the proposed pipeline. The architecture is intentionally designed to be modular, allowing each detection component to operate independently, thereby enabling straightforward updates and scalability. Furthermore, the system maintains a balance between performance and interpretability, making it suitable for practical applications such as digital forensics, media verification, and misinformation detection. Future work will focus on improving cross-modal fusion strategies and enhancing robustness against increasingly sophisticated deepfake generation techniques.

REFERENCES

[1] Hany Farid, "Digital image forensics," *Scientific American*, vol. 298, no. 6, pp. 66–71, 2008.

[2] Siwei Lyu, "Deepfake detection: Current challenges and next steps," *IEEE Signal Processing Magazine*, vol. 37, no. 2, pp. 26–35, 2020.

[3] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "FaceForensics++: Learning to detect manipulated facial images," *IEEE International Conference on Computer Vision*, 2019.

[4] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In Ictu Oculi: Exposing AI-created fake videos by detecting eye blinking," *IEEE International Workshop on Information Forensics and Security*, 2018.

[5] Pavel Korshunov and Sébastien Marcel, "Deepfakes: A new threat to face recognition?," *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

[6] David Güera and Edward J. Delp, "Deepfake video detection using recurrent neural networks," *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2018.

[7] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, "Detection of GAN-generated fake images over social networks," *IEEE Conference on Multimedia Information Processing and Retrieval*, 2018.

[8] Ning Yu, Larry S. Davis, and Mario Fritz, "Attributing fake images to GANs," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[9] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[10] Ramcharan Ramanaharan, Deepani B. Guruge, and Johnson I. Agbinya, "Deepfake video detection: Insights into model generalisation," *Data and Information Management*, 2025.

[11] Taiba Majid Wani and Irene Amerini, "Deepfake audio detection leveraging audio spectrogram and convolutional neural networks," *Springer Lecture Notes in Computer Science*, 2023.

[12] Aman Parikh, Kristen Pereira, Pranav Kumar, and Kailas Devadkar, "Audio-visual deepfake detection using multimodal deep learning," *International Conference on Intelligent Technologies*, 2023.

[13] Rineesh Babu P. and Madhu S. Nair, "Deepfake detection using multi-path CNN and convolutional attention mechanism," *IEEE MysuruCon*, 2022.

[14] Reshma Sunil, Parita Mer, Anjali Diwan, Rajesh Mahadeva, and Anuj Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey," *Heliyon*, vol. 11, 2025.

[15] M. D. Sarfaraz Momin, Abu Sufian, Debaditya Barman, Marco Leo, Cosimo Distante, and Naser Damer, "Explainable deepfake detection across different modalities," *Image and Vision Computing*, 2025.

[16] Leo Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[18] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.

[20] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1660–1664, 2018.

[21] Suneeta Satpathy, Álvaro Rocha, Sachi Nandan Mohanty, and Tanupriya Choudhury, *Intelligent Data-Driven Systems with Innovations in Artificial Intelligence*, CRC Press, 2025.

[22] Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, and Dhirendra Kumar Shukla, *Recent Trends in Intelligent Computing and Communication*, CRC Press, 2025.

[23] Dhirendra Kumar Shukla, Shabir Ali, and Sandhya Sharma, *Artificial Intelligence and Sustainable Innovation – Volume 2*, CRC Press, 2026.

[24] Amit Kumar Singh, Stefano Berretti, Ashima Anand, and Amrit Kumar Agrawal, *Digital Image Security: Techniques and Applications*, CRC Press, 2024.

[25] Natasa Kleanthous and Abir Hussain, *Machine Learning in Farm Animal Behavior Using Python*, CRC Press, 2025.

[26] Oluwafemi Ayotunde Oke and Nadire Cavus, "Electrocardiogram image classification using deep learning," *Iran Journal of Computer Science*, 2025.

[27] Shuqin Zhang, Meijing Liu, and Xiaogang Wang, "Bioelectronic medicine in rehabilitation," *Engineering Medicine*, 2026.

[28] Mitra Alakananda, *Machine Learning Methods for Data Quality Aspects in Edge Computing Platforms*, University of North Texas, 2023.

[30] Tilakachuri Balakrishna, B. Narendra, Mooray Harika Reddy, and Damarapati Jayasri, "Diagnosis of chronic kidney disease using random forest classification technique," Helix, vol. 7, no. 1, pp. 873–877, 2017.

[31] Tilakachuri Balakrishna, J. R. Annam, and D. Haritha, "Comparative analysis on liver benchmark datasets and prediction using supervised learning techniques," Indonesian Journal of Electrical Engineering and Computer Science, vol. 36, no. 2, 2024.