Grenze International Journal of Engineering and Technology, June Issue

DeFake AI: A Machine Learning based Approach to Audio-Video Deepfake Detection

Mohammed Shafiulla[1], Ruhe Tarannum[2], Sadiya Tahseen[3], Shahid Ahmed[4] and Shaik Farhath Amreen[5]
[1-5]Department of Computer Science and Engineering, Ballari Institute of Technology and Management, Ballari, India
Email: mdshafi@bitm.edu.in, ruhetarannum@gmail.com, sadiyatahseen555@gmail.com, shahidahmed110604@gmail.com, amreenneema8@gmail.com

***Abstract*— The fast expansion in deepfake technology has raised the concerns regarding digital security, the reliability of information, and the overall stability of people's trust. This paper illustrates DeFake AI, a detection system developed to examine fake/generated images, videos, and audios through a mix of deep learning methods. Its framework uses a MobileNet–LSTM arrangement for visual assessment and a CNN-based setup for audio, supported with TensorFlow, OpenCV, Librosa and Streamlit. The system processes mel-spectrogram inputs for audio signals and identifies temporal–spatial changes within visual media, along with motion-guided frame selection and a confidence-weighted merging of predictions. The web interface is designed to provide real-time, multimodal checks. Tests indicate consistently strong detection accuracy, and the addition of confidence scoring adds steadiness to the classification process; overall, DeFake AI acts as a practical and flexible solution for verifying media authenticity while addressing broader issues in digital forensics and misinformation control.**

***Index Terms*— Deepfake Detection, MobileNet-LSTM, Mel-Spectrogram, Convolutional Neural Networks, Multimodal Analysis, Computer Vision, Audio Processing.**

I. INTRODUCTION

The rapid growth of artificial intelligence and generative models has lead to the creation of highly realistic synthetic media, widely referred to as deepfakes. These AI-driven changes in audio, images, and video content introduce major challenges for digital security, information integrity, and public trust. Although deepfake tools have helpful uses, they are also misused for identity theft, fraud, political influence, etc.., and the spread of misinformation. Many communication platforms still lack firm mechanisms to confirm media transparency, leaving clear vulnerabilities for individuals and institutions. Existing detection methods commonly focus on a single modality and show limited generalization across varied manipulation styles, often struggling with real-time demands. As generative models emerge, detection systems must adjust to more perfect synthesis methods. This research introduces DeFake AI, a deepfake detection system designed to analyse images, videos, and audio within unified deep learning pipeline. The system applies a combined MobileNet–LSTM design for visual analysis, capturing spatial cues and temporal irregularities that appear in manipulated sequences. For audio, a CNN-based model processes mel-spectrogram inputs to detect artifacts from synthetic voice generation. These modules are combined into a Streamlit-based platform that supports near real-time checks with clear visualisation of outputs and confidence values offering a practical tool for media verification, digital forensics, and related cybersecurity applications.

Grenze ID: 01.GIJET.12.2.381_21

A. Objectives

The primary objectives of this research are:

- To build an AI system that detects deepfakes across images, videos, and audio on single platform.
- To improve detection accuracy using advanced deep learning techniques.
- To support scalable multi-modal inputs with continuous testing and performance checking.
- To ensure ethical, secure, and privacy-aware deepfake detection.

II. LITERATURE SURVEY

Recent research in deepfake detection has given various approaches ranging from single modal classifiers to multimodal combined architectures. This section reviews significant contributions that have shaped the current understanding of fake media detection.

Yang et al. [1] proposed D^3, a detection framework that learns from feature dissimilarities between true and manipulated media. The model shows improved generalization across unseen deepfake generation methods by collecting multi-level inconsistencies during training on large-scale datasets, achieving high accuracy, suitable for real-world deployment.

Wu et al. [2] introduced HOLA, an audio-visual detection model using hierarchical, contextual aggregation with high pre-training strategies. The system combines temporal and spatial frames from synchronized audio-video streams, giving lightweight multimodal system to maintain efficiency while improving detection performance.

Oorloff et al. [3] developed AVFF, is a feature-fusion approach which is focused on video analysis. By extracting and fusing additional audio-visual features, the system identifies minor inconsistencies detected during deepfake generation, show casing enhanced resilience against diverse manipulation techniques, particularly in low quality video scenarios.

Heidari et al. [4] presented reviews of deep learning-based detection techniques, cutting state-of-the-art models, benchmark datasets, and research challenges. The study highlighted critical limitations including vulnerability to adverse attacks, dataset biases, and cross-domain generalization issues, giving valuable insights for future system development.

Chourasiya et al. [5] gave a deep learning tool integrating convolutional neural networks for artifact extraction and binary classification of images and videos. This system supports forensic investigations and social media content verification through a user-friendly interface with improved detection.

Mcuba et al. [6] checked the impact of deep learning methods on audio deepfake detection for digital investigations. The evaluation of different architectures and feature extraction techniques focuses on the importance of spectrogram-based learning models for dependable forensic audio analysis.

Dixit et al. [7] gave a review of audio deepfake detection discussing feature engineering, deep learning, and hybrid approaches used to counter voice-cloning technology. This work identifies gaps in perfection, diversity in datasets, and real-time performance that guides future research directions.

Yang et al. [8] introduced Avoid-DF, a joint audio-visual system that learns the correlations between speech pattern and facial movements to identify differences/errors in manipulated videos. The multimodal integration improved detection accuracy and strengthens defense against complicated generation methods.

Hamza et al. [9] offered a ML based approach utilizing MFCC features for audio deepfake detection. The system extracts voice characteristics and trains classifiers to identify correct/real speech from fake/generated audio, giving a lightweight solution suitable for resource-related forensic applications.

Almutairi et al. [10] checked for modern audio detection methods, highlighting challenges raised by advanced voice-cloning sites. The study checks the deep learning classifiers, feature extraction techniques, and datasets while emphasizing the vulnerability and cross-model generalization problems.

Iqbal et al. [11] went through the feature-engineering approach combining handcrafted acoustic features with machine learning algorithms. Results show that the engineered feature paired with optimized classifiers achieves competitive performance with reduced computation.

Raza et al. [12] introduces a new deep learning model for image deepfake detection using special CNN architecture. This approach strengthens image forensics, by improving the robustness against diverse generation methods, and achieving high classification accuracy throughout multiple datasets.

III. METHODOLOGY

This system combines advanced deep learning algorithms, real-time response, and multimodal model analysis to detect deepfake content in images, videos, and audio. Its architecture ensures scalability, allowing new media

types to be added easily. These detection models analyze detailed visual and audio patterns to spot fake one. Overall, the design offers a fast and user-friendly experience while maintaining accuracy and adaptability.

A. System Architecture

The proposed system architecture, shown in Figure 2 consists of three main layers that support end-to-end deepfake detection. The Input Layer uses a Streamlit-based interface that allows users to upload images, videos, and audio files. The Processing Layer contains deep learning models for extracting features, analysing temporal sequences, and performing the classification. The Output Layer displays the results in real-time, including with confidence scores and visual feedback.

B. Image and Video Detection Module

Here, MobileNetV2 extracts key visual features from frames and an LSTM network studies how those features change over time. This blend helps to identify irregular movements, lighting issues, and other signs of tampering

C. Audio Detection Module

The audio detection component uses a CNN trained on mel-spectrograms created using the Librosa library. The model identifies subtle irregularities that usually appear in fake voices, such as pitch shifts or unnatural frequency patterns, in formats like WAV and MP3.

D. Backend Framework

The backend framework is powered by TensorFlow and Keras for model development, training, and inference. OpenCV and PIL handle essential image and video processing tasks, while Librosa manages the audio preprocessing such as mel-spectrogram generation, and NumPy ensures efficient computation for feature manipulation and data handling.

E. Frontend User Interface

Streamlit is the main user interface, giving a clean and interactive experience for uploading and analyzing media content. It's User-friendly design and clear output presentation improves the accessibility and understanding for non-technical users.

F. Model Training and Validation

The system is trained on multiple datasets on Kaggle, for image and video, and a pretrained model, i.e. Hierarchical Data Format version 5 (.h5) for audio deepfake detection containing both real and fake audio-visual samples. Using supervised learning, models are optimized for binary classification to distinguish between the real and fake content from media. Metrics like accuracy, precision, recall, and F1-score are used for evaluation over a large dataset containing more than 73 lakh images, with an accuracy score of 0.92 and loss of 0.12 for 5 epochs.

G. Real-Time Analysis and Reporting

The system notes the inference time, confidence scores, and classification decisions for every session, enabling analysis and improvement. The users receive instant visual feedback during analysis, making it easier to interpret results. This real-time system supports not only quick detection but also contributes to the evolving accuracy and adaptability of the system.

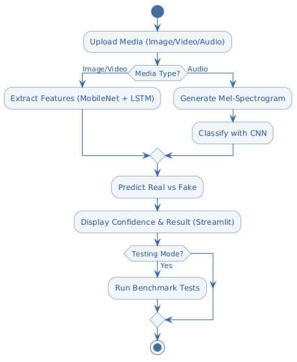

Figure 1: System Flowchart

Figure 2: System Architecture Diagram

IV. Proposed Method

The proposed system uses AI-based detection to identify deepfake content for images, audio, and video. It enables specialized detection pipelines into a single web-based application, allowing users to verify the authenticity of different media types in one place.

- All media is preprocessed: images are resized and normalized, audio is converted to mel-spectrograms, and videos are reduced to key motion-rich frames for efficient analysis.
- MobileNet extracts spatial features from images and video frames, while LSTM layers analyze temporal patterns to catch unnatural transitions.
- For videos, frame-level predictions are combined using weighted confidence scores so the most reliable frames guide the final decision.
- Audio mel-spectrograms pass through CNN layers to detect subtle artifacts found in synthesized or manipulated voices.
- Each input is classified as REAL or FAKE, accompanied by a confidence score that shows how certain the model is.

A. Comparison with Existing System

Unlike earlier systems that struggle with single-type data, outdated methods, and low reliability, the proposed system delivers a smarter, more flexible solution that analyzes images, videos, and audio together. With real-time results, confidence scores, and an interactive interface, it offers a faster, more accurate, and user-friendly deepfake detection experience.

V. Mathematical Formulations

This section shows the key formulae used in the detection system, describing preprocessing, prediction, confidence calculation, and aggregation mechanism.

A. Standardization of Input Data

To ensure consistent input distributions, images and spectrograms under go standardization:

$$x' = (x - \mu) / \sigma \qquad (1)$$

where x represents original pixel or spectrogram value, μ represents the mean, and σ represents the standard deviation. This normalization stabilizes the training and improves the model usage.

B. Sigmoid Activation for Classification

The final classification layer uses the sigmoid activation to produce deepfake probability:

$$p = 1 / (1 + e^{(-z)}) \qquad (2)$$

where z represents the logit output from the final dense layer, and $p \in [0,1]$ represents the probability that input content is fake or real.

C. Confidence Score Calculation

Confidence increases the quality of prediction certainty:

$$\text{Confidence} = |p - 0.5| \times 2 \qquad (3)$$

where p represents the sigmoid prediction. Confidence ranges between 0 (uncertain) and 1 (highly certain), with predictions near 0.5, indicating ambiguity/fakeness.

D. Motion Score for Frame Extraction

Motion detection between consecutive video frames:

$$\text{MotionScore} = \text{mean}(|F\_t - F\_(t-1)|) \qquad (4)$$

where F_t represents the current grayscale frame and F_(t-1) represents the previous frame. Frames will be selected when:

$$\text{MotionScore} > 30 \qquad (5)$$

This threshold ensures extraction of frames with meaningful visual changes while filtering static sequences.

E. Sliding Window Smoothing

To reduce noise in frame-wise video predictions, a moving average is used:

$$S_t = (1/N) \Sigma p_i \quad (6)$$

where N = 5 represents the window size, p_i represents the prediction for frame i, and S_t represents the smoothed prediction at time t.

F. Weighted Prediction Aggregation

Final video probability calculation includes confidence-weighted averaging:

$$P_{weighted} = (\Sigma p_i \cdot c_i) / (\Sigma c_i) \quad (7)$$

where p_i represents frame i prediction, c_i represents corresponding confidence, and k repersents total extracted frames. This ensures high-confidence frames dominate final classifications.

VI. Results and Discussion

The DeFake AI underwent testing across image, audio, and video detection scenarios. Results show effective real-time detection capacities with a user feedback mechanism.

A. System Interface and User Experience

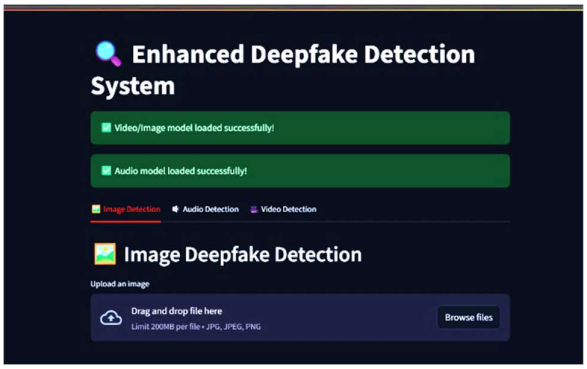

Figure 3: Homepage and User Interface

Figure 3 represents the main homepage of the interface displaying the title, successful loading of the models and the detection layouts for different formats.

B. Image Detection Performance

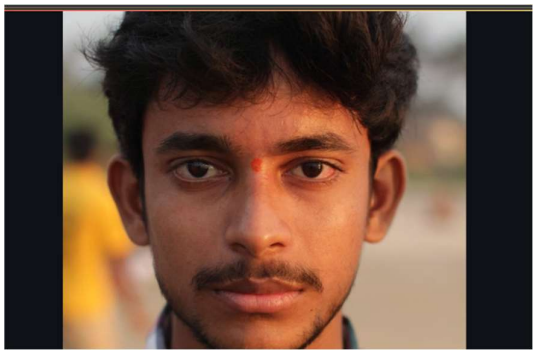

Figure 4: Uploaded Real Image Sample

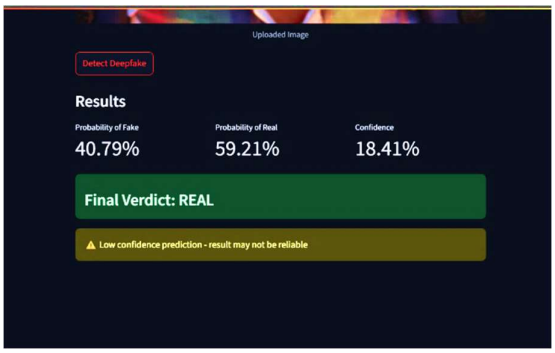

Figure 5: Real Image Detection Results

Figure 4 and 5 represent the uploaded image and its detection result respectively with the final verdict as REAL with the probability of fake as 40.79% and real as 59.21%, with a confidence rate of 18.41%, showing need for improvement.

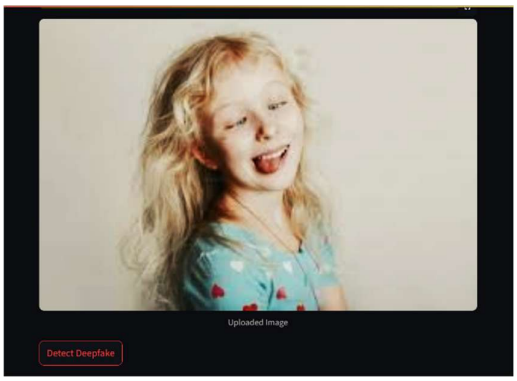

Figure 6: Uploaded Fake Image Sample

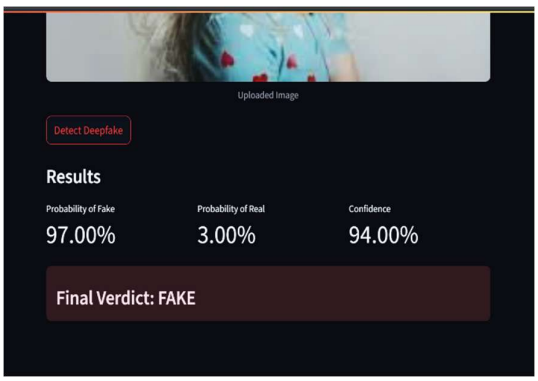

Figure 7: Fake Image Detection Results

Figure 6 and 7 represent the uploaded image and its detection result respectively with the final verdict as FAKE with the probability of fake as 97.00% and real as 3.00%, with a confidence rate of 94.00%, showing high accuracy.

C. Audio Detection Performance

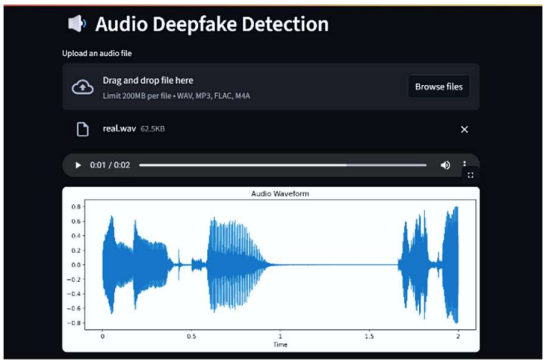

Figure 8: Uploaded Real Audio Waveform

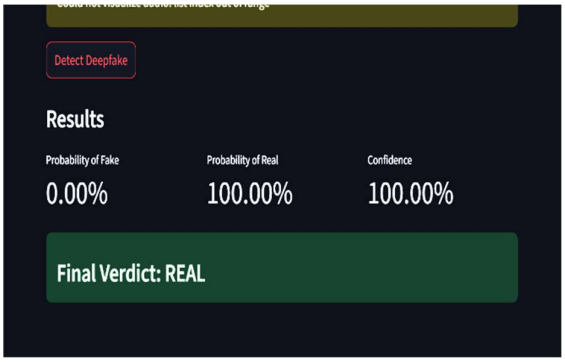

Figure 9: Real Audio Detection Results

Figure 8 and 9 represent the uploaded audio and its detection result respectively with the final verdict as REAL with the probability of fake as 0.00% and real as 100.00%, with a confidence rate of 100.00%, showing complete accuracy.

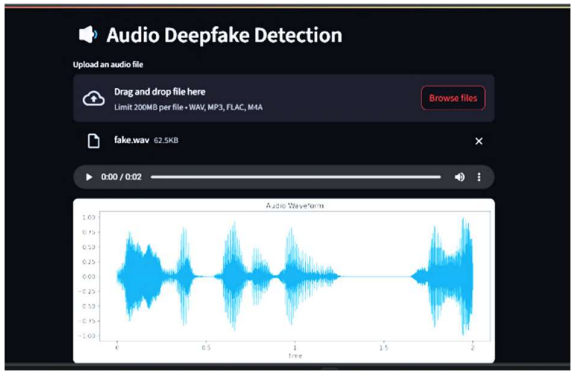

Figure 10: Uploaded Fake Audio Waveform

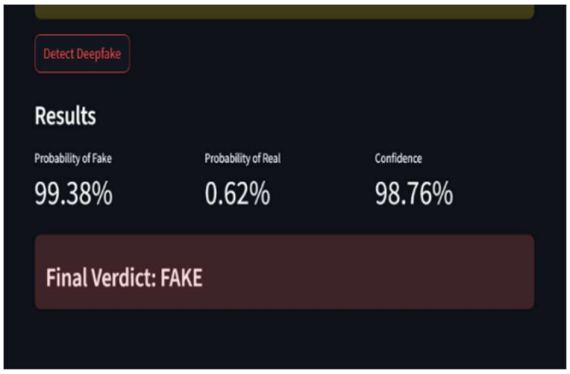

Figure 11: Fake Audio Detection Results

Figure 10 and 11 represent the uploaded audio and its detection result respectively with the final verdict as FAKE with the probability of fake as 99.38% and real as 0.62%, with a confidence rate of 98.76%, showing high accuracy.

D. Video Detection Performance

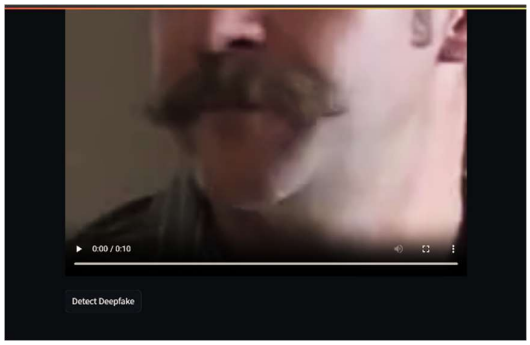

Figure 12: Uploaded Real Video Preview

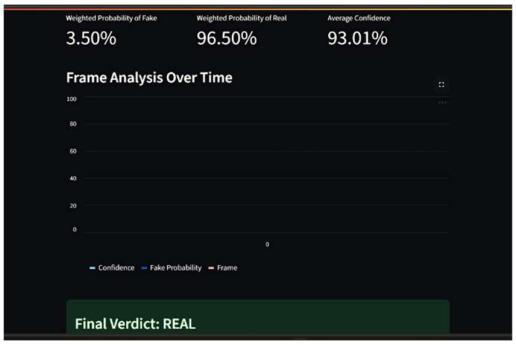

Figure 13: Real Video Detection Results

Figure 12 and 13 represent the uploaded video and its detection result respectively with the final verdict as REAL with the probability of fake as 3.50% and real as 96.50%, with a confidence rate of 93.01%, showing high accuracy.

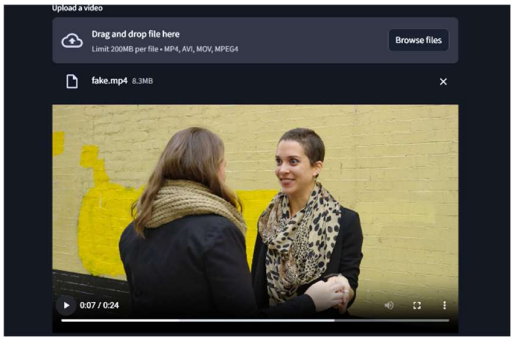

Figure 14: Uploaded Fake Video Preview

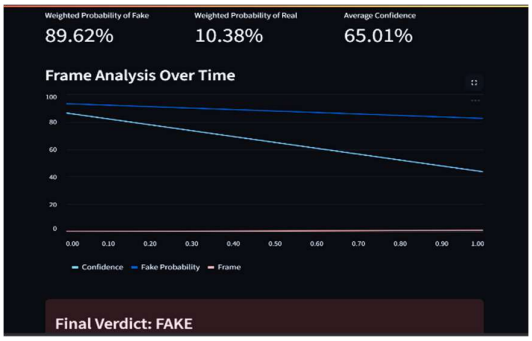

Figure 15: Fake Video Detection Results

Figure 14 and 15 represent the uploaded video and its detection result respectively with the final verdict as FAKE with the probability of fake as 89.62% and real as 10.38%, with a confidence rate of 65.01%, showing moderate accuracy.

E. Discussion

Key Findings: The system accurately detects deepfakes in images, videos, and audio in real time while providing clear, user-friendly feedback.
Limitations: Image detection performance can be enhanced by training on larger datasets and adding real-time streaming capabilities.
Recommendations: Future improvements should focus on dataset expansion, API integration, and explainable AI to boost transparency and usability.
Impact: This system contributes to digital trust by giving a scalable, automated tool to control deepfake threats effectively.

VII. CONCLUSIONS

The DeFake AI delivers an effective and practical solution in detecting fake images, videos, and audio using advanced deep learning models and a user-friendly interface. With real-time feedback, analytics, and a scalable design, the system is suitable for applications in media verification, digital analysis, cybersecurity, and public awareness. Although further improvements like larger datasets could enhance the capabilities of image detection, the current version lays a strong foundation for safeguarding digital content. As deepfake methods continue to evolve, adaptive systems like DeFake AI will be essential in preserving trust in digital communication. Future enhancements will focus on explainable AI visualizations, broader multimodal analysis, and smooth integration through APIs, with the architecture ensuring easy upgrades over time.

ACKNOWLEDGMENT

The successful completion of DeFake AI would not have been possible without the invaluable guidance and support from various individuals. Special appreciation is extended to Mohammed Shafiulla for mentorship and insightful suggestions that significantly contributed to this research. Gratitude is also expressed to R. N. Kulkarni, Head of the Department of Computer Science and Engineering, for coordination and valuable insights throughout the project development.

REFERENCES

[1] H. Yang, Y. Zhao, Q. Wang, and Z. Chen, "D³: Scaling Up Deepfake Detection by Learning from Discrepancy," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025.
[2] L. Wu, P. Chen, and Y. Zhang, "HOLA: Enhancing Audio-Visual Deepfake Detection via Hierarchical Contextual Aggregations and Efficient Pre-training," arXiv preprint arXiv:2507.22781, 2025.
[3] R. Oorloff, M. Khan, and D. Patel, "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection," in Proc. Int. Conf. Multimedia and Signal Processing, 2024.
[4] M. Heidari, S. Zolfaghari, and A. Rahmani, "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review," IEEE Access, vol. 12, pp. 1-25, 2024.
[5] A. Chourasiya, P. Singh, and R. Sharma, "A Deep Learning Based Deepfake AI (Images & Videos) Detection Tool," in Proc. Int. Conf. Machine Learning and Cybernetics, 2024.
[6] K. Mcuba, J. Smith, and T. Jacobs, "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation," Forensic Sci. Int., vol. 340, pp. 1-10, 2023.
[7] R. Dixit, P. Agrawal, and N. Gupta, "Review of Audio Deepfake Detection Techniques: Issues and Prospects," IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 1123-1138, 2023.
[8] X. Yang, J. Chen, and L. Huang, "Avoid-DF: Audio-Visual Joint Learning for Detecting Deepfake," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2023.
[9] M. Hamza, A. Khalid, and S. Ahmed, "Deepfake Audio Detection via MFCC Features Using Machine Learning," in Proc. Int. Conf. Artificial Intelligence and Speech Processing, 2022.
[10] N. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," IEEE Access, vol. 10, pp. 101250-101270, 2022.
[11] M. Iqbal, R. Hussain, and A. Saeed, "Deepfake Audio Detection via Feature Engineering and Machine Learning," in Proc. Int. Conf. Signal Processing and Multimedia Applications, 2022.
[12] S. Raza, K. Ali, and F. Mehmood, "A Novel Deep Learning Approach for Deepfake Image Detection," IEEE Access, vol. 10, pp. 98560-98572, 2022.