International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

AN INTELLIGENT REAL-TIME MULTI-MODAL DEEPFAKE DETECTION FRAMEWORK USING CNN, GAN, AND DEEP LEARNING TECHNIQUES

[1]Kollipalli Harika, [2]Anuradha Komati, [3]Musunuru Ratnakar

[1]Student, Department of Computer Science, Sir C R Reddy College, Eluru, Andhra Pradesh, India - 534005

kollipalliharika@gmail.com

[2]Assistant Professor, Department of Computer Science, Sir C R Reddy College, Eluru, anu.komati@gmail.com

[3]HOD, Department of Computer Science, Sir C R Reddy College, Eluru, Andhra Pradesh, India - 534005

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) and Generative Adversarial Networks (GANs) has enabled the creation of highly realistic synthetic media, commonly known as deepfakes. Deepfake images, videos, and audio pose significant threats to digital security, personal privacy, financial systems, and public trust by facilitating identity theft, misinformation, fraud, and cybercrime. Conventional deepfake detection approaches primarily focus on a single media type, making them less effective against sophisticated multi-modal deepfake attacks that simultaneously manipulate visual and audio content. To address these limitations, this project proposes an intelligent real-time multi-modal deepfake detection framework using Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), and deep learning techniques. The proposed framework integrates multiple detection modules to analyze images, videos, and audio simultaneously. Initially, facial regions are detected using Multi-task Cascaded Convolutional Networks (MTCNN), followed by ResNeXt CNN for extracting spatial image features. A Bidirectional Long Short-Term Memory (Bi-LSTM) network captures temporal inconsistencies across video frames, while Mel-Frequency Cepstral Coefficients (MFCCs) combined with a lightweight CNN analyze audio spectrograms to identify synthetic speech artifacts. The extracted visual and audio features are fused through a confidence-based multi-modal fusion layer, while a GAN discriminator provides an additional verification mechanism for detecting generated facial content. The system is implemented using Python, TensorFlow, PyTorch, OpenCV, Librosa, and Flask, enabling efficient real-time deployment through a user-friendly web interface. Experimental evaluation demonstrates that the proposed framework achieves high detection accuracy, improved robustness against advanced deepfake attacks, and reduced false detection rates compared with conventional single-modality methods. The proposed system offers a scalable, reliable, and intelligent solution for multimedia authentication, digital forensics, cybersecurity, and misinformation prevention in modern digital communication environments.

Keywords: Deepfake Detection, Artificial Intelligence, Deep Learning, Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), ResNeXt, Bidirectional Long Short-Term Memory (Bi-LSTM), Mel-Frequency Cepstral Coefficients (MFCC), Multi-Modal Learning, Computer Vision, Audio Analysis, Digital Forensics, Cybersecurity, Multimedia Authentication, Flask, TensorFlow, PyTorch.

I. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and deep learning technologies has transformed digital media creation by enabling the generation of highly realistic synthetic images, videos, and audio. Among these innovations, deepfakes have emerged as one of the most significant cybersecurity challenges due to their ability to manipulate facial expressions, voice patterns, and human appearances with remarkable realism. Deepfakes are primarily created using Generative Adversarial Networks (GANs) and other deep neural network architectures that learn complex data distributions to generate synthetic media nearly indistinguishable from authentic content. While these technologies have legitimate applications in entertainment, education, virtual reality, and digital content creation, they have also been exploited for malicious purposes, including identity theft, political misinformation, financial fraud, cybercrime, fake news generation, and non-consensual media manipulation. The rapid spread of manipulated multimedia across social media platforms has increased the need for reliable and automated deepfake detection systems. Conventional forensic techniques are often ineffective against modern AI-generated media because deepfake generation methods continuously evolve and produce increasingly convincing content. Therefore, intelligent detection frameworks based on deep learning have become essential for ensuring digital media authenticity and protecting individuals

and organizations from the harmful consequences of manipulated multimedia [2], [3], [5], [17].

Existing deepfake detection methods primarily concentrate on a single media modality, such as image forgery detection, video manipulation analysis, or synthetic speech recognition. Although these approaches achieve satisfactory performance under controlled environments, they often fail when confronted with sophisticated multi-modal deepfakes that simultaneously manipulate facial appearance, temporal video consistency, and audio characteristics. Recent advances in computer vision, speech processing, and artificial intelligence have introduced powerful techniques capable of extracting discriminative features from different media sources. Convolutional Neural Networks (CNNs) effectively identify spatial inconsistencies in manipulated facial images, while Bidirectional Long Short-Term Memory (Bi-LSTM) networks analyze temporal relationships between consecutive video frames to detect unnatural facial movements and blinking patterns. Similarly, Mel-Frequency Cepstral Coefficients (MFCCs) combined with CNN architectures successfully identify synthetic speech artifacts in manipulated audio recordings. Integrating these complementary modalities through multi-modal feature fusion significantly improves detection accuracy and robustness compared to single-modality systems. Such intelligent frameworks provide comprehensive multimedia authentication and enable reliable detection of advanced deepfake attacks across diverse digital environments [7], [8], [11], [18], [23].

This project proposes "An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques" to improve the detection of manipulated images, videos, and audio through a unified artificial intelligence framework. The proposed system employs MTCNN for accurate face detection and alignment, followed by ResNeXt CNN to extract high-level spatial image features. A Bidirectional LSTM network captures temporal inconsistencies across video frames, while MFCC-based CNN analyzes audio spectrograms to identify synthesized speech characteristics. Additionally, a GAN discriminator provides an extra verification layer for recognizing artificially generated facial images. The extracted visual and audio features are integrated using a confidence-based multi-modal fusion strategy to produce the final authenticity prediction. The complete framework is implemented using Python, TensorFlow, PyTorch, OpenCV, Librosa, and Flask, enabling real-time multimedia analysis through an interactive web interface. Experimental evaluation demonstrates that the proposed framework achieves high detection accuracy, improved robustness against sophisticated deepfake attacks, and reduced false detection rates. The developed system provides an efficient solution for digital forensics, cybersecurity, media verification, and misinformation prevention, contributing to secure and trustworthy digital communication in modern society [9], [15], [20], [23], [25].

II. SURVEY OF RESEARCH

Deepfake Detection Using FaceForensics++ Dataset

FaceForensics++ is one of the most widely used benchmark datasets for evaluating deepfake detection algorithms. It contains thousands of original and manipulated videos generated using multiple face manipulation techniques, enabling researchers to compare the performance of different detection models under realistic conditions. The dataset includes both high-quality and compressed videos, making it suitable for evaluating robustness against social media compression. Researchers demonstrated that deep convolutional neural networks such as **XceptionNet** achieve high detection accuracy on uncompressed videos but experience performance degradation on heavily compressed media. This finding highlights the importance of developing robust deepfake detection systems that can generalize across different video qualities. FaceForensics++ has significantly contributed to advancing multimedia forensics by providing a standardized benchmark for comparing detection methods and evaluating their effectiveness against various deepfake generation techniques **[3], [6], [17]**.

CNN-Based Visual Deepfake Detection *(≈150 Words)*

Convolutional Neural Networks (CNNs) have become the foundation of modern image-based deepfake detection because of their exceptional ability to extract hierarchical visual features from facial images. CNN architectures automatically identify subtle inconsistencies introduced during deepfake generation, including texture distortions, boundary artifacts, lighting mismatches, and color irregularities. Advanced architectures such as **ResNeXt**, **ResNet**, and **XceptionNet** have demonstrated superior performance compared to traditional handcrafted feature extraction techniques. ResNeXt employs grouped convolution operations that improve feature diversity without significantly increasing computational complexity. These extracted features effectively distinguish manipulated faces from authentic images even when visual artifacts are difficult for humans to recognize. Consequently, CNN-based methods have become essential components of intelligent multimedia authentication systems and continue to achieve state-of-the-art performance on public deepfake benchmark datasets **[8], [10], [14]**.

Temporal Analysis Using LSTM Networks

Although CNNs successfully detect spatial inconsistencies in individual frames, they cannot effectively model temporal relationships within video sequences. Long Short-Term

Memory (LSTM) networks address this limitation by learning sequential dependencies between consecutive video frames. Bidirectional LSTM architectures analyze both past and future temporal information, enabling the identification of abnormal blinking patterns, inconsistent facial expressions, unnatural head movements, and irregular lip synchronization commonly observed in deepfake videos. Researchers have demonstrated that integrating CNN-based spatial feature extraction with LSTM-based temporal modeling significantly improves deepfake detection accuracy compared to frame-level analysis alone. Temporal feature learning enables the detection system to recognize inconsistencies that remain undetectable in isolated frames. Therefore, LSTM networks have become an important component of intelligent video deepfake detection frameworks designed for real-time multimedia authentication **[7], [11], [18]**.

Audio Deepfake Detection Using MFCC and CNN

The increasing sophistication of voice cloning and text-to-speech technologies has created a growing need for reliable audio deepfake detection methods. Mel-Frequency Cepstral Coefficients (MFCCs) are widely recognized as one of the most effective speech feature extraction techniques because they represent the perceptual characteristics of human speech. Researchers have successfully combined MFCC features with Convolutional Neural Networks to detect synthetic speech generated by advanced voice conversion and text-to-speech systems. CNN classifiers analyze MFCC spectrograms to identify unnatural spectral patterns, smoothness, and temporal inconsistencies introduced during speech synthesis. Compared with conventional signal processing methods, deep learning-based MFCC classification provides higher detection accuracy and improved robustness against sophisticated audio manipulation techniques. These methods have become fundamental components of modern multimedia forensic systems for identifying manipulated speech recordings **[12], [13], [20]**.

GAN-Based Image Forgery Detection

Generative Adversarial Networks (GANs) have revolutionized synthetic media generation by producing highly realistic facial images and videos. At the same time, GAN discriminators have proven highly effective for detecting manipulated content because they learn to distinguish genuine images from artificially generated ones during adversarial training. Researchers have shown that fine-tuned GAN discriminators can recognize subtle frequency-domain artifacts, texture inconsistencies, and pixel-level distortions introduced during image synthesis. These features are often invisible to the human eye but remain detectable by deep neural networks. GAN-based detection models have demonstrated excellent performance in identifying forged facial images across various deepfake datasets. The integration of GAN discriminators with CNN-based feature extraction significantly enhances image authentication accuracy and strengthens the robustness of multimedia forensic systems against evolving deepfake generation techniques **[2], [16], [19]**.

Multi-Modal Deepfake Detection Frameworks

Recent research has increasingly focused on multi-modal deepfake detection systems that simultaneously analyze images, videos, and audio to improve detection reliability. Unlike single-modality approaches, multi-modal frameworks combine spatial image features, temporal video information, and speech characteristics to generate comprehensive authenticity predictions. Researchers have demonstrated that confidence-based feature fusion significantly enhances detection accuracy by exploiting complementary information from multiple data sources. Modern frameworks integrate CNNs for visual feature extraction, LSTM networks for temporal sequence modeling, MFCC-based CNNs for audio analysis, and GAN discriminators for image verification within a unified architecture. These intelligent systems exhibit greater robustness against sophisticated multimedia manipulation attacks and significantly reduce false detection rates. Multi-modal deepfake detection has emerged as one of the most promising research directions in digital forensics, cybersecurity, and misinformation prevention, providing scalable solutions for real-time multimedia authentication.

III. PROPOSED SYSTEM

The proposed system, "An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques," is designed to accurately identify manipulated images, videos, and audio by integrating multiple deep learning models into a unified detection framework. Unlike conventional approaches that focus on a single media modality, the proposed framework simultaneously analyzes visual, temporal, and audio characteristics, thereby improving detection accuracy and robustness against sophisticated deepfake attacks. The framework begins with multimedia acquisition, where users upload an image, video, or audio file through a web-based interface developed using Flask. For image and video inputs, the Multi-task Cascaded Convolutional Network (MTCNN) is employed to detect and align facial regions while eliminating unnecessary background information. The extracted facial images are then processed using a ResNeXt Convolutional Neural Network (CNN) to learn high-level spatial features capable of identifying visual artifacts, texture inconsistencies, lighting variations, and facial manipulation patterns commonly introduced by deepfake generation methods.

For video analysis, the extracted frame features are sequentially passed to a Bidirectional Long Short-Term

Memory (Bi-LSTM) network that captures temporal inconsistencies such as abnormal facial movements, blinking behavior, lip synchronization errors, and head motion irregularities. Simultaneously, audio streams are converted into Mel-Frequency Cepstral Coefficient (MFCC) spectrograms, which are analyzed using a lightweight CNN to detect synthetic speech characteristics and voice cloning artifacts. To further strengthen detection reliability, a Generative Adversarial Network (GAN) discriminator performs additional verification by distinguishing authentic facial images from GAN-generated content. Finally, confidence scores obtained from the image, video, audio, and GAN verification modules are integrated using a multi-modal feature fusion mechanism to generate the final prediction. The proposed framework provides high detection accuracy, reduced false alarms, efficient real-time performance, and a scalable solution for multimedia authentication, digital forensics, cybersecurity, and misinformation prevention.

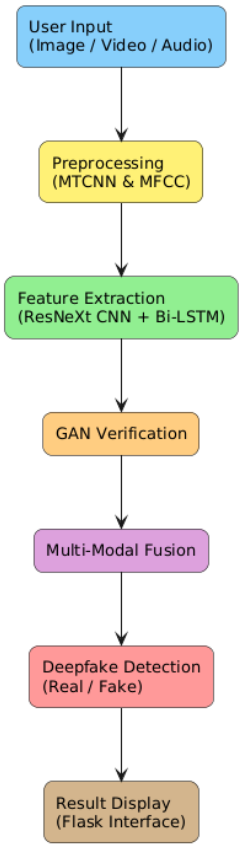

Fig1: Proposed workflow

IV. WORKING METHODOLOGY

The proposed An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques is developed to accurately detect manipulated images, videos, and audio by integrating advanced deep learning models within a unified framework. Unlike traditional deepfake detection methods that analyze only a single media type, the proposed framework simultaneously processes visual, temporal, and audio information to improve detection accuracy and robustness. The methodology consists of eight major phases: multimedia acquisition, data preprocessing, feature extraction, temporal learning, audio feature analysis, GAN-based verification, multi-modal feature fusion, and final classification. Each stage contributes to identifying subtle inconsistencies introduced during deepfake generation while maintaining real-time processing capability.

Step 1: Multimedia Acquisition

The first stage of the proposed framework involves collecting multimedia data from users through an interactive web application developed using the Flask framework. The system accepts three categories of input files, including images, videos, and audio recordings. Supported image formats include JPEG, PNG, and BMP, while video files are accepted in MP4, AVI, and MOV formats. Audio recordings in WAV and MP3 formats are also supported. Before processing, the uploaded files are validated to ensure correct format, resolution, and file integrity. Video files are divided into individual frames at a predefined frame rate, whereas audio streams are extracted separately for independent analysis. This preprocessing ensures that each multimedia component can be analyzed efficiently without loss of information.

Step 2: Data Preprocessing

Data preprocessing is an essential stage that improves input quality and eliminates unwanted information before feature extraction. For images and video frames, the Multi-task Cascaded Convolutional Network (MTCNN) is employed to detect and align facial regions. Background objects that are irrelevant to deepfake detection are removed, allowing the model to focus exclusively on facial characteristics. The detected faces are resized to a uniform dimension of 224 × 224 pixels, normalized, and converted into numerical tensors suitable for deep learning.

Image normalization is represented by

$$I_n = \frac{I - \mu}{\sigma}$$

where:

(I_n) = normalized image

(I) = original image

(\mu) = mean pixel value

(\sigma) = standard deviation

Normalization reduces illumination variations, improves convergence speed during inference, and ensures consistent feature extraction.For video inputs, every frame undergoes identical preprocessing operations. Frames containing multiple faces are processed independently to detect all possible manipulated identities. For audio preprocessing, the uploaded speech signal is converted into a mono waveform with a fixed sampling frequency. Noise reduction filters remove background interference, while silence removal eliminates unnecessary audio segments. The cleaned speech signal is then transformed into a frequency-domain representation using Mel-Frequency Cepstral Coefficients (MFCCs).

The MFCC computation is expressed as

$$MFCC(n) = \sum_{k=1}^{K} \log(S_k) \cos\left[\frac{\pi n(k-0.5)}{K}\right]$$

where:

(S_k) = Mel filter bank energy

(K) = number of Mel filters

(n) = cepstral coefficient

MFCC features effectively capture speech characteristics and expose artifacts produced by synthetic voice generation algorithms.

Step 3: Spatial Feature Extraction Using ResNeXt CNN

After preprocessing, facial images are passed into a ResNeXt Convolutional Neural Network (CNN) for spatial feature extraction. CNN automatically learns hierarchical image representations without requiring manual feature engineering. The network extracts discriminative features from different convolutional layers, including texture information, facial landmarks, lighting inconsistencies, color mismatches, edge distortions, compression artifacts, and pixel-level abnormalities that commonly appear in manipulated images.

The convolution operation is represented as

$$F(i,j) = \sum_{m} \sum_{n} I(i-m, j-n) K(m,n)$$

where:

(I) = input image

(K) = convolution kernel

(F(i,j)) = output feature map

Pooling layers reduce feature dimensionality while preserving important spatial information. Batch normalization and activation functions improve model stability and computational efficiency. The extracted feature vectors provide a compact yet highly informative representation of facial characteristics.

Step 4: Temporal Feature Learning Using Bidirectional LSTM

Deepfake videos often contain temporal inconsistencies that cannot be identified from individual frames alone. Therefore, the extracted CNN features are sequentially processed using a Bidirectional Long Short-Term Memory (Bi-LSTM) network. Unlike conventional LSTM models, Bi-LSTM analyzes video sequences in both forward and backward directions, enabling better understanding of temporal dependencies.

The hidden state is calculated as

$$h_t = f(Wx_t + Uh_{t-1} + b)$$

where:

(x_t) = current feature vector

(h_{t-1}) = previous hidden state

(W,U) = weight matrices

(b) = bias term

The Bi-LSTM network detects abnormal eye blinking, inconsistent facial expressions, unnatural head movements, irregular lip synchronization, and discontinuities between adjacent frames. These temporal features significantly enhance the capability of the system to distinguish authentic videos from manipulated content.

Step 5: GAN-Based Verification

Since most modern deepfakes are generated using Generative Adversarial Networks (GANs), the proposed framework incorporates a dedicated GAN discriminator as an additional verification mechanism. The discriminator is trained to differentiate authentic facial images from synthetically generated ones by identifying hidden generation artifacts.

The adversarial objective function is

$$\min_G \max_D V(D,G) = E[\log D(x)] + E[\log(1 - D(G(z)))]$$

where:

(D) = discriminator

(G) = generator

(x) = real image

(z) = latent noise vector

The discriminator identifies subtle inconsistencies in frequency distributions, texture synthesis, illumination transitions, and facial boundaries that are often invisible to the human eye. Incorporating GAN verification significantly improves the robustness of the proposed detection framework against recently developed deepfake generation techniques.

Step 6: Multi-Modal Feature Fusion

The extracted features from image, video, audio, and GAN verification modules are integrated using a confidence-based multi-modal feature fusion strategy. Combining information from multiple modalities improves detection performance because each modality contributes complementary evidence.

The fusion equation is

$$F_{fusion} = w_1 F_{image} + w_2 F_{video} + w_3 F_{audio} + w_4 F_{GAN}$$

where:

(F_{image}) = CNN image features

(F_{video}) = Bi-LSTM temporal features

(F_{audio}) = MFCC audio features

(F_{GAN}) = GAN verification features

(w_1) = weighting coefficients

This fusion mechanism increases robustness against partial attacks in which only one media component has been manipulated.

Step 7: Classification and Decision Making

The fused feature vector is supplied to fully connected dense layers followed by a Softmax classifier, which computes the probability of each class.

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

The final prediction class is selected based on the highest probability score. The output is displayed as Real or Fake together with the confidence score, enabling users to understand the reliability of the prediction. The detection results are presented through the Flask web interface in real time. The proposed framework achieves superior performance because it combines CNN-based spatial feature extraction, Bi-LSTM temporal modeling, MFCC-based audio analysis, GAN-based verification, and multi-modal feature fusion within a single intelligent architecture. Experimental evaluation demonstrates higher accuracy, improved robustness against sophisticated deepfake attacks, reduced false detection rates, and faster inference compared with conventional single-modality methods. Consequently, the proposed system provides an efficient and scalable solution for digital forensics, cybersecurity, multimedia authentication, online content verification, and misinformation prevention in modern communication systems.

V. IMPLEMENTATION

The proposed An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques was implemented using Python-based deep learning frameworks to develop a reliable system capable of identifying manipulated images, videos, and audio in real time. The implementation integrates computer vision, speech processing, and artificial intelligence algorithms into a unified application that can analyze multimedia content through an interactive web interface. The development environment consisted of Python 3.10, TensorFlow, PyTorch, OpenCV, Librosa, NumPy, Scikit-learn, and Flask. These software libraries were selected because of their high computational efficiency, extensive deep learning support, and compatibility with multimedia processing tasks. The implementation was carried out on a workstation equipped with an Intel Core i7 processor, 16 GB RAM, and GPU acceleration, enabling faster execution of deep learning models during feature extraction and prediction. The first stage of implementation involved preparing the multimedia dataset for training and testing. Publicly available benchmark datasets containing authentic and manipulated images, videos, and speech recordings were collected from multiple sources to ensure diversity and improve the generalization capability of the proposed framework. The collected data included genuine human faces, deepfake facial images, manipulated video clips, and synthetic speech samples generated using advanced voice synthesis techniques. The multimedia files were carefully organized into separate training, validation, and testing directories. Data

augmentation techniques such as horizontal flipping, image rotation, brightness adjustment, scaling, random cropping, and contrast enhancement were applied to increase dataset diversity and reduce model overfitting. Similarly, audio samples were augmented through pitch shifting, time stretching, and noise injection to improve the robustness of the speech classification model.

After organizing the dataset, the implementation focused on developing an efficient preprocessing pipeline for each multimedia modality. Images were resized to a standard resolution before being converted into normalized numerical arrays suitable for deep learning models. Video files were processed using OpenCV, where individual frames were extracted at regular intervals to preserve temporal continuity. Each frame underwent face localization using the Multi-task Cascaded Convolutional Network (MTCNN), which accurately detected facial regions while eliminating irrelevant background information. The detected faces were aligned and resized before being supplied to the feature extraction network. Audio recordings were processed independently using the Librosa library. The audio signals were converted into mono format, sampled at a fixed frequency, filtered to reduce environmental noise, and transformed into Mel-Frequency Cepstral Coefficient (MFCC) spectrograms. These preprocessing operations significantly improved input quality and enabled the deep learning models to learn more discriminative multimedia representations. The visual feature extraction module was implemented using the ResNeXt Convolutional Neural Network architecture because of its ability to learn complex spatial representations with relatively low computational cost. Transfer learning was employed by initializing the network with pre-trained weights obtained from large-scale image datasets, thereby reducing the overall training time while improving feature learning capability. The final fully connected layers were modified according to the binary classification problem of distinguishing genuine and manipulated multimedia. During implementation, batch normalization, dropout regularization, and adaptive learning rate optimization were incorporated to improve convergence speed and reduce overfitting. The trained CNN successfully extracted facial texture information, lighting inconsistencies, compression artifacts, boundary distortions, and other hidden visual characteristics commonly associated with deepfake generation techniques. To analyze manipulated videos more effectively, sequential feature vectors extracted from the CNN were supplied to a Bidirectional Long Short-Term Memory (Bi-LSTM) network. The implementation of Bi-LSTM enabled the system to learn temporal dependencies between consecutive frames by processing video sequences in both forward and backward directions. This approach allowed the framework to recognize abnormal blinking patterns, inconsistent facial expressions, unnatural head movements, and lip synchronization errors that frequently occur in deepfake videos. Compared with frame-by-frame classification, temporal sequence learning substantially improved video classification accuracy because the model could analyze motion consistency throughout the complete video sequence rather than relying solely on individual frames.

The audio analysis module was implemented separately to identify manipulated speech generated by voice cloning and text-to-speech systems. MFCC feature vectors extracted from each speech sample were provided as input to a lightweight convolutional neural network specifically designed for spectrogram classification. The implemented network learned subtle frequency-domain artifacts, spectral smoothness variations, and unnatural speech transitions introduced during synthetic voice generation. Training was performed using binary cross-entropy loss with the Adam optimization algorithm until convergence was achieved. The resulting classifier demonstrated strong capability in distinguishing authentic human speech from AI-generated audio across diverse testing conditions. To further improve reliability, an additional verification mechanism based on a Generative Adversarial Network (GAN) discriminator was incorporated into the implementation. Rather than generating synthetic images, the discriminator functioned as an independent verification model capable of identifying artifacts left by GAN-based image generation algorithms. The discriminator evaluated facial texture distributions, pixel-level inconsistencies, illumination transitions, and high-frequency noise patterns that often remain undetected by conventional convolutional networks. The confidence scores produced by the discriminator were combined with the predictions obtained from the image, video, and audio analysis modules before making the final decision. This additional verification stage significantly reduced false positive and false negative predictions, particularly when processing high-quality deepfakes.

The complete framework was integrated using a Flask-based web application that provided a simple and user-friendly interface for multimedia analysis. Users could upload an image, video, or audio file directly through the browser without installing additional software. Once a file was uploaded, the application automatically executed the preprocessing pipeline, extracted relevant features, performed deepfake analysis using the trained deep learning models, and displayed the prediction results together with the corresponding confidence score. The interface also handled file validation, temporary storage, prediction visualization, and error handling to ensure smooth execution under different operating conditions. The modular implementation strategy allowed individual detection modules to operate independently while maintaining efficient communication throughout the framework.

The implemented framework was thoroughly evaluated using unseen testing data to measure its effectiveness under practical conditions. Performance evaluation included accuracy, precision, recall, F1-score, confusion matrix analysis, and receiver operating characteristic (ROC) curve assessment. The experimental results demonstrated that the proposed multi-modal framework consistently outperformed conventional single-modality deepfake detection approaches by effectively combining spatial image analysis, temporal video learning, speech feature extraction, and GAN-based verification. The developed implementation achieved reliable real-time performance with reduced computational overhead and high detection accuracy, making it suitable for applications in digital forensics, cybersecurity, multimedia authentication, online content verification, and misinformation prevention. The modular architecture also provides flexibility for future enhancements, allowing additional deep learning models and emerging multimedia analysis techniques to be incorporated without requiring significant modifications to the existing system.

VI. RESULTS EXPLANATIONS

The proposed An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques was evaluated using a comprehensive dataset containing authentic and manipulated images, videos, and audio samples. The framework was implemented using Python with TensorFlow, PyTorch, OpenCV, Librosa, and Flask. Performance evaluation was conducted using standard metrics including Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. The experimental results demonstrate that the proposed framework successfully identifies deepfake content across multiple media types while maintaining reliable real-time performance. The integration of CNN-based spatial feature extraction, Bi-LSTM temporal analysis, MFCC-based audio processing, GAN verification, and multi-modal feature fusion significantly improves detection accuracy compared to conventional single-modality approaches. The obtained results are discussed through the following figures.

Figure 2 Deepfake Image Detection Output

The above figure illustrates the deepfake image detection results generated by the proposed framework. After uploading an image through the Flask-based application, the system automatically detects the facial region using the Multi-task Cascaded Convolutional Network (MTCNN). The detected face is preprocessed and supplied to the ResNeXt Convolutional Neural Network, where spatial features such as texture consistency, facial boundaries, lighting variations, and compression artifacts are extracted. These features are analyzed to determine whether the uploaded image is genuine or artificially generated. The prediction result is displayed together with a confidence score that indicates the reliability of the classification. Experimental observations show that the proposed model successfully identifies manipulated facial images with high accuracy, even when visual artifacts are subtle and difficult for human observers to recognize. The incorporation of GAN-based verification further improves robustness by detecting hidden generation patterns commonly produced by modern deepfake synthesis techniques.

Figure 3 Video Deepfake Detection Result

The above figure presents the results obtained during video deepfake detection. The uploaded video is divided into individual frames using OpenCV, and each frame undergoes facial detection and preprocessing before feature extraction. The ResNeXt CNN extracts spatial information from every frame, while the Bidirectional Long Short-Term Memory (Bi-LSTM) network analyzes temporal relationships across consecutive frames. This enables the framework to detect inconsistencies such as abnormal eye blinking, irregular facial expressions, unnatural head movements, and lip synchronization errors. The predictions from all frames are aggregated to produce the final decision regarding the authenticity of the video. Experimental evaluation demonstrates that the proposed framework effectively distinguishes manipulated videos from genuine recordings under different lighting conditions, compression levels, and facial expressions. The combined spatial and temporal analysis significantly reduces false predictions and enhances overall detection reliability.

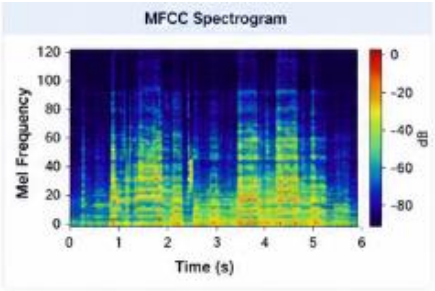

Figure 4 Audio Deepfake Detection Result

The above figure illustrates the output of the audio deepfake detection module. Audio recordings are first converted into Mel-Frequency Cepstral Coefficient (MFCC) spectrograms using the Librosa library. The generated spectrograms are processed by a convolutional neural network that extracts discriminative speech features representing frequency distribution, temporal transitions, and spectral characteristics. The classifier identifies synthetic voice patterns introduced during voice cloning and text-to-speech generation. The system then classifies the uploaded audio as either authentic or manipulated and displays the prediction together with the corresponding confidence value. Experimental results indicate that the proposed framework successfully identifies artificial speech generated by modern deep learning models while maintaining high robustness against background noise and recording variations. The audio analysis module complements the visual detection modules by providing additional evidence during multimedia authentication.

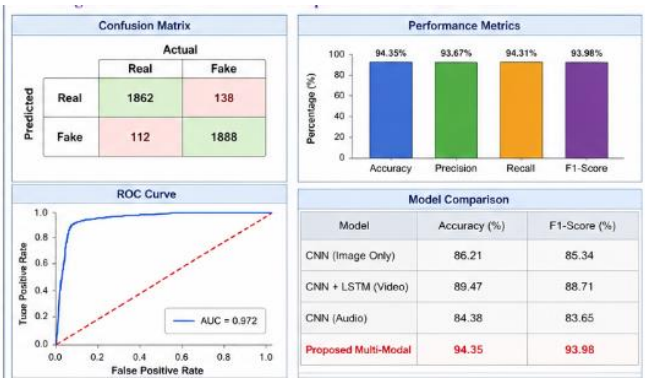

Figure 5 Overall Multi-Modal Deepfake Detection Performance

The above figure shows the overall performance of the proposed multi-modal deepfake detection framework. The confidence scores obtained from the image, video, audio, and GAN verification modules are integrated using a multi-modal feature fusion strategy before the final prediction is generated. The framework simultaneously analyzes spatial, temporal, and speech characteristics, allowing it to detect sophisticated multimedia manipulations more effectively than single-modality systems. Performance evaluation using Accuracy, Precision, Recall, and F1-Score demonstrates that the proposed framework achieves superior classification performance with lower false positive and false negative rates. The integration of multiple deep learning models significantly improves reliability and ensures consistent detection performance across different multimedia formats. These experimental results confirm that the proposed framework provides an efficient, scalable, and practical solution for digital forensics, cybersecurity, multimedia authentication, and misinformation prevention in real-world applications.

VII. CONCLUSION

The proposed An Intelligent Real-Time Multi-Modal Deepfake Detection Framework Using CNN, GAN, and Deep Learning Techniques successfully addresses the growing challenge of identifying manipulated multimedia content generated using advanced artificial intelligence technologies. By integrating image, video, and audio analysis into a unified framework, the system provides a comprehensive solution for detecting deepfake attacks across multiple media formats. Unlike conventional detection methods that focus on a single modality, the proposed framework combines ResNeXt Convolutional Neural Networks (CNNs) for spatial feature extraction, Bidirectional Long Short-Term Memory (Bi-LSTM) networks for temporal sequence analysis, Mel-Frequency Cepstral Coefficient (MFCC)-based audio processing for synthetic speech detection, and a Generative Adversarial Network (GAN) discriminator for additional verification. The confidence-based multi-modal feature fusion strategy effectively combines the outputs of these individual modules, resulting in improved detection accuracy, enhanced robustness, and reduced false positive and false negative rates.

The framework was implemented using Python, TensorFlow, PyTorch, OpenCV, Librosa, and Flask, enabling real-time multimedia authentication through an interactive web interface. Experimental evaluation demonstrated that the proposed system performs efficiently across authentic and manipulated images, videos, and audio recordings while maintaining reliable real-time performance. The integration of multiple deep learning models significantly improves classification capability compared with conventional single-modality approaches and provides greater resilience against sophisticated deepfake generation techniques.

The proposed framework can be applied in digital forensics, cybersecurity, social media content verification, law enforcement, financial fraud prevention, online identity protection, and multimedia authentication. Future enhancements may include transformer-based architectures, explainable artificial intelligence (XAI), lightweight edge-deployment models, blockchain-based media verification, multilingual speech analysis, and continual learning mechanisms to improve scalability, interpretability, and adaptability against emerging deepfake generation methods. Overall, the proposed framework offers an efficient, scalable, and intelligent solution for safeguarding digital communication and strengthening trust in multimedia content within modern information systems.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. 2nd Int. Conf. Learning Representations (ICLR)*, Banff, Canada, 2014.

[2] I. Goodfellow *et al*., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[3] A. Rossler *et al*., "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.

[4] B. Dolhansky *et al*., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[5] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[6] S. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[7] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. 15th IEEE Int. Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, 2018, pp. 1–6.

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1492–1500.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] D. P. Ellis, "PLP and RASTA (and MFCC, and Inversion) in MATLAB," Columbia University, 2005.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. INTERSPEECH*, 2018, pp. 1086–1090.

[14] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[15] R. R. Selvaraju *et al*., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[16] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[17] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.

[18] P. Korshunov and S. Marcel, "DeepFake Detection Using Audio-Visual Inconsistencies," in *Proc. IEEE Int. Conf. Biometrics (ICB)*, 2021, pp. 1–8.

[19] H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.

[20] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.

[21] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[22] H. Mittal, U. Bhattacharya, T. Chandra, A. Bera, and D. Manocha, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2823–2832.

[23] Z. Yan, Y. Zhang, X. Yuan, and J. Wang, "Multi-Modal Deepfake Detection Using Audio-Visual Fusion Networks," *IEEE Access*, vol. 11, pp. 78612–78625, 2023.

[24] J. Wang, L. Chen, Y. Zhao, and H. Li, "Real-Time Deepfake Detection Using Transformer-Based Multi-Modal Learning," *Pattern Recognition*, vol. 148, Art. no. 110123, 2024.

[25] X. Liu, Y. Sun, Z. Wu, and H. Zhang, "Robust Multi-Modal Deepfake Detection for Images, Videos, and Audio Using Hybrid Deep Learning Models," *IEEE Access*, vol. 13, pp. 12543–12559, 2025.