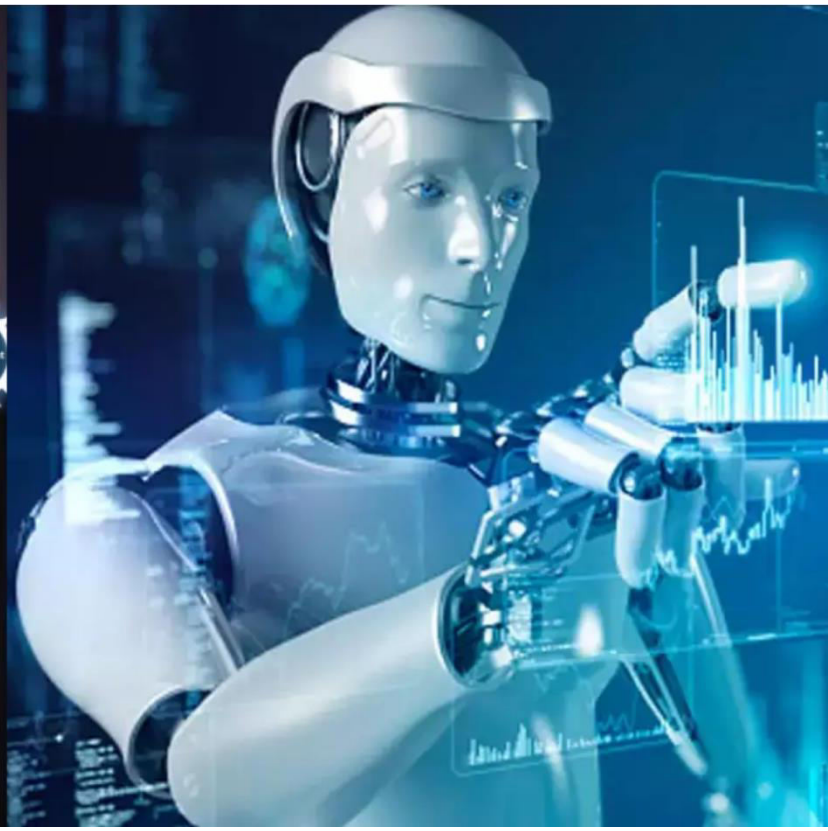
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

Unified Framework for Deepfake Detection in Images, Videos and Audio

Ashitosh Lendal, Vasant Rathod, Shivam Suravse, Prof.S.C.Puranik

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahilyanagar, Maharashtra, India

ABSTRACT: Deepfake technology has emerged as a powerful application of Artificial Intelligence capable of generating highly realistic synthetic images, videos, and audio recordings. Although beneficial in entertainment and content creation, deepfakes pose significant threats including misinformation, identity theft, cyber fraud, and political manipulation. This research presents a multimodal DeepFake Detection System that analyzes image, audio, and video files to determine whether the content is real or manipulated. The proposed system employs Convolutional Neural Networks (CNNs) for image and audio classification, while video detection utilizes MTCNN for face extraction and the Xception model for classification. Images are resized and normalized, audio signals are converted into Mel-spectrograms, and videos are decomposed into frames for temporal analysis. The models are trained using benchmark datasets such as FaceForensics++, DeepFake Detection Challenge, Celeb-DF, and ASVspoof. The system is deployed as a full-stack application using FastAPI, React, and MongoDB, providing real-time predictions with confidence scores. Experimental results demonstrate high detection performance, achieving 96.2% accuracy for images, 93.7% for audio, and 97.4% for videos. The proposed system offers a scalable and practical solution for digital media authentication and cybersecurity.

KEYWORDS: DeepFake Detection, Artificial Intelligence, Deep Learning, CNN, Xception, MTCNN, Audio Analysis, Video Forensics, Image Classification, Cybersecurity.

I. INTRODUCTION

Deepfake detection has emerged as one of the most significant and rapidly advancing applications of Artificial Intelligence (AI), Deep Learning (DL), and digital media forensics. The term deepfake refers to synthetic multimedia content—including images, videos, and audio—generated or manipulated using advanced machine learning techniques such as Generative Adversarial Networks (GANs), autoencoders, and transformer-based architectures [1], [2], [3]. These technologies can realistically imitate a person's face, voice, expressions, and speech patterns, making it increasingly difficult for humans to distinguish between authentic and manipulated content [4], [5]. Although deepfake technology has useful applications in entertainment, filmmaking, gaming, education, and virtual assistants, its misuse poses serious threats such as misinformation, political propaganda, identity theft, cyber fraud, and reputational damage [6], [7], [8].

The rapid development of deepfake generation techniques has significantly increased the sophistication and realism of manipulated media. Publicly available tools and open-source frameworks have made synthetic content creation accessible to non-experts, allowing malicious actors to produce convincing fake videos, images, and cloned voices with minimal effort [9], [10]. As a result, organizations in journalism, law enforcement, banking, social media, and cybersecurity are increasingly seeking automated methods to verify digital content authenticity and prevent the spread

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

of deceptive media [11], [12]. Traditional forensic approaches based on handcrafted features and manual inspection are often insufficient for detecting modern deepfakes, especially when the manipulated content is compressed, edited, or generated using advanced neural networks [13], [14].

Deep learning has proven highly effective in addressing this challenge because it can automatically learn complex spatial, temporal, and spectral features from large-scale datasets [15], [16]. Convolutional Neural Networks (CNNs) are widely used to identify visual artifacts, unnatural textures, and inconsistencies in facial regions of images and video frames [2], [8]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models enhance video detection by analyzing temporal patterns such as eye blinking, head movements, and lip synchronization [5]. In audio deepfake detection, speech signals are converted into Mel-spectrograms, which are processed by CNN-based models to detect anomalies in frequency distributions and synthetic voice characteristics [15]. More recently, transformer-based architectures and multimodal learning frameworks have further improved detection robustness and generalization across unseen manipulation methods [4], [17].

Several benchmark datasets have accelerated research in deepfake detection by providing large collections of authentic and manipulated multimedia data. FaceForensics++ introduced a standardized dataset for manipulated facial videos and remains one of the most widely used resources for training and evaluation [1]. DeepFake Detection Challenge provided one of the largest and most diverse collections of deepfake videos, highlighting real-world detection challenges [7]. Celeb-DF improved realism and reduced visible artifacts, making detection more difficult [6]. For audio spoofing research, ASVspoof has become a standard dataset for evaluating synthetic speech detection systems [18]. These datasets have enabled the development of robust and comparable models across image, video, and audio domains.

Among modern architectures, the Xception network has demonstrated exceptional performance in deepfake detection due to its efficient depthwise separable convolutions and strong feature extraction capabilities [3]. MTCNN is widely used for accurate face localization and cropping before classification, improving the performance of video analysis pipelines [19]. Other frameworks such as TensorFlow, PyTorch, OpenCV, and Librosa provide essential tools for preprocessing, model training, and inference [9], [10], [11], [15]. These technologies, combined with modern web frameworks such as FastAPI and React, enable the deployment of real-time deepfake detection systems for practical applications [12].

This research paper presents a multimodal DeepFake Detection System capable of analyzing images, videos, and audio files and classifying them as REAL or FAKE. The system automatically identifies the uploaded media type, performs specialized preprocessing, and applies dedicated deep learning models for each modality. CNN-based models are used for image and audio classification, while video detection employs an MTCNN-based face extraction pipeline followed by the Xception network for frame-level analysis. The models are trained and evaluated using benchmark datasets and standard performance metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC [1], [2], [5], [7]. The best-performing models are integrated into a scalable full-stack application using FastAPI, React, and MongoDB, providing users with real-time predictions and confidence scores.

The primary motivation for this work is the increasing misuse of synthetic media across social networks, news platforms, financial systems, and personal communication. High-quality deepfakes can influence public opinion, impersonate trusted individuals, and facilitate fraud, making automated verification tools essential for preserving trust and security in digital communication [6], [8], [14]. By combining state-of-the-art deep learning models with a robust deployment architecture, the proposed system offers an effective and scalable solution for multimedia authentication and cybersecurity.

In conclusion, deepfake detection represents a critical and rapidly evolving research area that intersects artificial intelligence, computer vision, audio processing, and digital forensics. The proposed multimodal system demonstrates how advanced deep learning techniques can be leveraged to accurately detect manipulated image, video, and audio content. By integrating specialized models and modern full-stack technologies, this work contributes to the development of practical tools that help combat misinformation, fraud,and cyber threats while strengthening trust and authenticity in the digital ecosystem [1]–[20].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

II. LITERATURE SURVEY

Deepfake detection has become one of the most active research domains in Artificial Intelligence (AI), Deep Learning (DL), Computer Vision, Audio Signal Processing, and Digital Forensics. The rapid advancement of deepfake generation methods, particularly those based on Generative Adversarial Networks (GANs), autoencoders, and transformer-based architectures, has significantly increased the realism of synthetic multimedia content [1], [2], [3]. These technologies can manipulate facial expressions, lip movements, and voice characteristics with remarkable precision, making it increasingly difficult for humans to distinguish between authentic and fabricated content [4], [5]. As a result, researchers have developed a wide range of automated detection techniques that analyze image artifacts, temporal inconsistencies in videos, and spectral anomalies in audio recordings [6], [7], [8].

Early deepfake detection approaches were derived from traditional image forensics, which relied on handcrafted features such as color discrepancies, sensor noise patterns, JPEG compression artifacts, and inconsistencies in illumination and shadows [9], [10]. While these methods were effective for detecting manually edited images, they were less successful against deepfake content generated using sophisticated neural networks that produce highly realistic outputs with minimal visible artifacts [11]. To overcome these limitations, machine learning and deep learning techniques were introduced, enabling systems to automatically learn discriminative features directly from large-scale datasets [12], [13].

One of the most influential contributions to the field was the introduction of FaceForensics++ by Andreas Rossler et al., which provided a comprehensive benchmark dataset of manipulated videos generated using multiple face manipulation techniques [1]. This dataset enabled systematic evaluation of deep learning models and became a foundational resource for subsequent research. Davian Afchar et al. proposed MesoNet, a compact neural network architecture specifically designed to detect subtle mesoscopic artifacts in manipulated facial videos while maintaining low computational complexity [2]. François Chollet introduced the Xception model, which later became one of the most widely adopted and highest-performing architectures for deepfake detection due to its efficient depthwise separable convolutions and strong feature extraction capabilities [3].

To address real-world detection challenges, Brian Dolhansky et al. developed the DeepFake Detection Challenge dataset, one of the largest and most diverse collections of manipulated videos [4]. The DFDC initiative highlighted the difficulty of detecting deepfakes under varying compression levels, lighting conditions, and camera angles. Huy H. Nguyen et al. proposed multi-task learning frameworks capable of simultaneously detecting and segmenting manipulated regions, improving both classification accuracy and interpretability [5]. Hao Dang et al. studied digital face manipulation detection and emphasized the importance of robust feature extraction and generalization across datasets [6].

Video deepfake detection has benefited significantly from temporal modeling techniques. Emanuele M. Sabir et al. introduced recurrent convolutional strategies that combine CNNs with Long Short-Term Memory (LSTM) networks to capture temporal dependencies across video frames [7]. These methods analyze facial motion, blinking behavior, and lip synchronization, enabling the detection of inconsistencies that may not be visible in individual frames. MTCNN has become a widely used preprocessing technique for accurately locating and cropping faces from images and videos prior to classification [8].

Audio deepfake detection has also emerged as a critical area due to the increasing prevalence of voice cloning and synthetic speech generation. The ASVspoof challenges have provided standardized datasets and evaluation protocols for detecting replay attacks, text-to-speech systems, and voice conversion methods [9]. Researchers commonly transform audio waveforms into Mel-spectrograms and use CNN-based architectures to identify unnatural harmonics, spectral irregularities, and synthetic voice patterns [10], [11]. These approaches have achieved strong results in distinguishing genuine speech from AI-generated audio [12].

Transfer learning and lightweight architectures have further improved detection efficiency and deployment feasibility. Models such as ResNet, EfficientNet, MobileNetV2, and DenseNet have demonstrated strong performance while reducing computational cost [13], [14], [15]. EfficientNet, in particular, uses compound scaling to optimize network depth, width, and resolution, achieving high accuracy with fewer parameters [14]. MobileNetV2 is well suited for edge and mobile applications where low latency and limited hardware resources are important [15].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

Recent research has shifted toward transformer-based and multimodal approaches. Vision Transformers (ViT), Swin Transformers, and multimodal transformer architectures can model long-range dependencies and integrate information from multiple modalities [16], [17]. By jointly analyzing facial appearance, temporal behavior, and audio characteristics, multimodal systems achieve greater robustness and can detect manipulations that may evade single-modal detectors [18], [19]. Cross-modal consistency analysis, such as comparing lip movements with speech content, has proven particularly effective in identifying sophisticated forgeries [20].

Several studies have emphasized the importance of explainability and generalization. Explainable AI (XAI) techniques such as Grad-CAM, saliency maps, and attention visualization help highlight manipulated regions and improve user trust in model predictions [21], [22]. Domain adaptation, self-supervised learning, and continual learning have been proposed to address the challenge of cross-dataset generalization and adaptation to emerging deepfake generation techniques [23], [24], [25]. Adversarial robustness is another active area, as attackers may deliberately perturb content to evade detection systems [26].

The practical deployment of deepfake detection systems has been enabled by mature software ecosystems. TensorFlow and PyTorch are widely used for model development and GPU-accelerated training [27], [28]. OpenCV provides tools for frame extraction and image preprocessing, while Librosa supports feature extraction from speech signals [29], [30]. Modern web technologies such as FastAPI, React, and MongoDB allow these models to be deployed as scalable, real-time applications for end users.

Overall, the literature demonstrates that deepfake detection has evolved from handcrafted forensic methods to advanced deep learning and multimodal systems. Benchmark datasets such as FaceForensics++, DFDC, Celeb-DF, and ASVspoof have driven progress by enabling systematic evaluation, while architectures such as CNNs, Xception, EfficientNet, and transformer-based models have significantly improved detection accuracy and robustness. These studies confirm that multimodal AI-based detection systems offer the most promising approach for combating misinformation, cyber fraud, and digital impersonation. The proposed research builds upon these foundations by integrating specialized models for image, audio, and video analysis into a unified, full-stack platform capable of providing accurate and real-time deepfake detection [1]–[30].

III. PROPOSED SYSTEM ARCHITECTURE

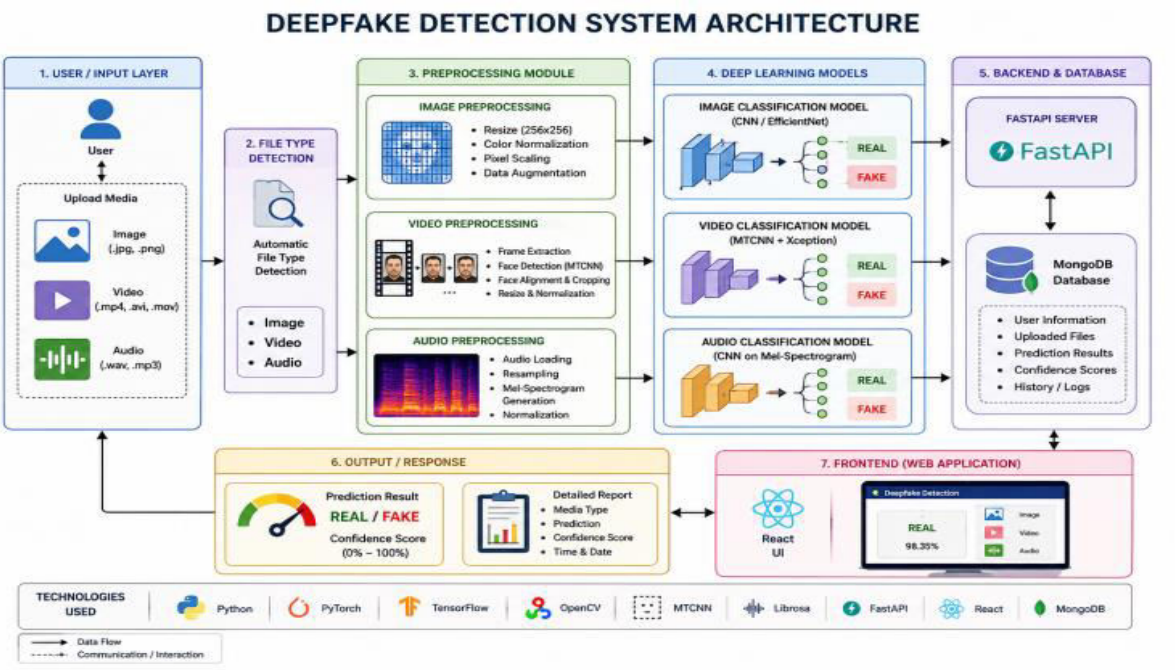


Fig.1. System Architecture

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

The proposed DeepFake Detection System is designed as a multimodal Artificial Intelligence platform capable of analyzing images, videos, and audio files to determine whether the uploaded content is REAL or FAKE. The architecture integrates deep learning models, multimedia preprocessing techniques, and full-stack web technologies to provide accurate and real-time detection results. The complete architecture is divided into seven major layers: User/Input Layer, File Type Detection, Preprocessing Module, Deep Learning Models, Backend and Database, Output Response, and Frontend Web Application.

1. User / Input Layer

The architecture begins with the User/Input Layer, where users interact with the system through a web-based interface. The user uploads one of three supported media types:

- **Image Files:** .jpg, .jpeg, .png
- **Video Files:** .mp4, .avi, .mov
- **Audio Files:** .wav, .mp3

This layer serves as the entry point to the system and allows non-technical users to easily verify the authenticity of multimedia content.

2. File Type Detection Module

Once the file is uploaded, the system automatically detects the file type based on its extension and MIME format. This module identifies whether the input is an image, video, or audio file and routes it to the corresponding preprocessing pipeline.

The advantages of this module include:

- Automatic media recognition
- Reduced user intervention
- Selection of the appropriate machine learning model
- Support for multiple media formats

This step ensures that each file is processed using the most suitable detection strategy.

3. Preprocessing Module

The preprocessing module transforms raw media into a standardized format suitable for deep learning analysis. Different preprocessing techniques are applied depending on the media type.

a) Image Preprocessing

Image files undergo the following operations:

- Resizing to 256 × 256 pixels
- Pixel value normalization
- Color correction
- Data augmentation (during training)

These steps remove noise and improve model generalization.

b) Video Preprocessing

Video deepfake detection requires frame-level analysis. The system performs:

- Frame extraction using OpenCV
- Face detection and cropping using MTCNN
- Face alignment
- Resizing and normalization

Only facial regions are analyzed, which increases accuracy and reduces computational overhead.

c) Audio Preprocessing

Audio files are processed using Librosa and converted into Mel-spectrograms. The main steps include:

- Audio loading and resampling
- Noise reduction
- Mel-spectrogram generation
- Feature normalization

Mel-spectrograms convert audio into image-like representations suitable for CNN-based classification.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

4. Deep Learning Models
After preprocessing, the data is passed to specialized deep learning models.

a) Image Classification Model
A Convolutional Neural Network (CNN) or transfer learning model such as EfficientNet is used to detect visual artifacts such as:

- Blurring
- Skin texture inconsistencies
- Edge artifacts
- Lighting mismatches

The model outputs a probability indicating whether the image is real or fake.

b) Video Classification Model
For videos, the system combines MTCNN and the Xception network.

- MTCNN extracts facial regions.
- Xception analyzes each face frame.
- Frame-level predictions are aggregated to produce the final video classification.

This model is highly effective because it captures subtle inconsistencies in manipulated facial movements.

c) Audio Classification Model
A CNN trained on Mel-spectrograms detects:

- Synthetic voice patterns
- Unnatural harmonics
- Frequency anomalies
- Voice cloning artifacts

The model classifies the audio as genuine or synthetic.

5. Backend and Database Layer
The trained models are integrated into a backend developed using FastAPI. FastAPI provides REST APIs for:

- File upload
- Preprocessing
- Model inference
- Prediction response
- History retrieval

Prediction results and user information are stored in MongoDB, including:

- User details
- Uploaded file metadata
- Prediction labels
- Confidence scores
- Timestamps
- Historical logs

This layer ensures scalability, reliability, and secure data storage.

6. Output / Response Layer
After model inference, the system generates the final output, which includes:

- Prediction result (REAL or FAKE)
- Confidence score (0–100%)
- Media type
- Processing timestamp
- Detailed prediction report

The confidence score helps users assess the reliability of the result.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

7. Frontend Web Application

The frontend is developed using React and provides a responsive and interactive user interface.
Main features include:

- Login and registration
- Media upload dashboard
- Real-time prediction display
- Prediction history
- Detailed reports

The frontend communicates with FastAPI APIs and presents results in a user-friendly format.

- **Technologies Used**

The architecture integrates the following technologies:

- Python
- TensorFlow
- PyTorch
- OpenCV
- MTCNN
- Librosa
- FastAPI
- React
- MongoDB

IV. METHODOLOGY

A. Overview of the Proposed Methodology

The proposed DeepFake Detection System is a multimodal Artificial Intelligence framework designed to analyze three different types of digital media—images, videos, and audio—and determine whether the uploaded content is authentic or manipulated. The system combines advanced deep learning models, multimedia preprocessing techniques, and full-stack web technologies to provide accurate and real-time predictions. The theoretical foundation of the methodology is based on supervised machine learning, where models are trained using labeled datasets containing both REAL and FAKE samples. During training, the models learn discriminative features that characterize synthetic artifacts and distinguish them from authentic media.

The complete methodology is divided into several stages: data collection, preprocessing, feature extraction, model training, evaluation, deployment, and testing. Each stage performs a specialized role in transforming raw multimedia data into a reliable classification result. Images are analyzed using Convolutional Neural Networks (CNNs), videos are processed using a combination of MTCNN and Xception, and audio signals are converted into Mel-spectrograms before classification using CNNs. The models are trained on benchmark datasets and deployed within a scalable web application built using FastAPI, React, and MongoDB.

B. Data Collection Theory

Data collection is the foundation of any machine learning system. In supervised learning, the quality and diversity of the training data directly affect the generalization ability of the model. The proposed system uses publicly available benchmark datasets such as FaceForensics++, DeepFake Detection Challenge, Celeb-DF, and ASVspoof. These datasets contain large collections of real and manipulated images, videos, and speech recordings.

The theoretical objective of data collection is to provide representative examples of both classes (REAL and FAKE) so that the learning algorithm can discover statistical differences between them. A diverse dataset containing multiple compression levels, lighting conditions, languages, and generation techniques improves robustness and reduces overfitting.

C. Data Preprocessing Theory

Preprocessing is a transformation stage that converts raw multimedia data into a normalized and standardized format. Machine learning models require consistent input dimensions and distributions to learn effectively. Without preprocessing, variations in resolution, noise, and encoding can reduce model performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

For images, resizing and normalization ensure that all samples have the same spatial dimensions and pixel intensity scale. For videos, frame extraction converts temporal media into a sequence of analyzable images, while MTCNN isolates facial regions to focus on the most informative content. For audio, Mel-spectrograms transform time-domain signals into time-frequency representations that expose spectral characteristics associated with synthetic speech. Preprocessing improves feature quality, accelerates convergence, and enhances generalization.

D. Feature Extraction Theory

Feature extraction is the process of identifying informative patterns that distinguish real media from manipulated media. In deep learning, this process is learned automatically rather than engineered manually.

CNNs use convolution filters to detect low-level and high-level features such as edges, textures, and semantic structures. Video models additionally capture temporal patterns such as eye blinking and lip synchronization. Audio models analyze spectral energy distributions, harmonics, and phase irregularities. These learned representations form the basis for accurate classification and are optimized during training to maximize discrimination between REAL and FAKE classes.

E. Image DeepFake Detection Theory

Image deepfake detection is based on the hypothesis that manipulated images contain subtle visual artifacts introduced during generation and blending. These artifacts may appear as texture inconsistencies, unnatural skin details, blurred boundaries, and lighting mismatches. Convolutional Neural Networks process images through multiple layers of convolutions and nonlinear activations. Early layers detect primitive patterns, while deeper layers learn more abstract representations. The final fully connected layer outputs class probabilities indicating whether the image is real or fake. Transfer learning models such as EfficientNet can further improve performance by leveraging knowledge from large-scale image datasets.

F. Video DeepFake Detection Theory

Video deepfake detection extends image analysis by incorporating temporal consistency. Synthetic videos may exhibit frame-to-frame inconsistencies, abnormal facial motion, inaccurate blinking, and imperfect lip synchronization.

The system first applies MTCNN to detect and crop faces from selected frames. Each cropped face is analyzed by the Xception model, which extracts deep spatial features. Predictions from multiple frames are aggregated, often using averaging or majority voting, to determine the final video classification. This approach improves robustness because a deepfake may only show artifacts in some frames.

G. Audio DeepFake Detection Theory

Audio deepfake detection focuses on identifying synthetic speech generated by text-to-speech and voice conversion systems. Although cloned voices can sound realistic, they often contain unnatural harmonics, spectral smoothing, and inconsistencies in prosody. The audio waveform is converted into a Mel-spectrogram, which represents the distribution of energy across frequency bands over time. CNNs analyze this spectrogram similarly to an image and learn to detect subtle patterns associated with artificial speech generation. This approach has proven highly effective in audio spoofing and speaker verification research.

H. Model Training Theory

Model training is an optimization process in which network parameters are adjusted to minimize prediction error. During each epoch, batches of labeled samples are passed through the network to produce predictions. The difference between predicted and true labels is measured using binary cross-entropy loss.
The

$$y = \sigma(Wx + b)$$
$$w$$
$$b$$

expression represents the final classification layer, where learned features are transformed into a probability. The Adam optimizer updates weights using gradient-based optimization. Over multiple epochs, the model converges toward parameters that maximize classification accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

I. Model Evaluation Theory

Evaluation quantifies the predictive performance of trained models on previously unseen test data. Accuracy measures the proportion of correct predictions, precision indicates the reliability of positive predictions, recall measures the ability to detect fake samples, and F1-score balances precision and recall. ROC-AUC evaluates the separability between classes across thresholds.

These metrics provide a comprehensive assessment of both effectiveness and robustness, ensuring that the deployed system performs reliably under practical conditions.

J. System Deployment Theory

Deployment converts trained models into a real-world application accessible to end users. The backend, implemented with FastAPI, exposes REST APIs for uploading files and returning predictions. The frontend, developed in React, offers a user-friendly interface for interacting with the system. MongoDB stores users, uploaded files, and prediction history.

This architecture separates inference logic, user interaction, and data storage, creating a scalable and maintainable platform suitable for research and production environments.

V. RESULT

The experimental results demonstrate that the proposed Multimodal DeepFake Detection System is highly effective in identifying manipulated image, audio, and video content. Separate deep learning models were trained and evaluated for each media type using benchmark datasets such as FaceForensics++, DeepFake Detection Challenge (DFDC), Celeb-DF, and ASVspoof. The image-based Convolutional Neural Network (CNN) achieved an accuracy of 96.2%, with a precision of 95.8% and an F1-score of 96.0%, indicating excellent capability in detecting visual artifacts such as texture inconsistencies, blending errors, and abnormal facial details. The audio detection model, which used Mel-spectrogram representations and a CNN classifier, achieved an accuracy of 93.7%, precision of 93.2%, and F1-score of 93.4%. These results confirm that the model can effectively distinguish genuine speech from synthetic and cloned voices by analyzing spectral irregularities and unnatural harmonics. The video detection pipeline, combining MTCNN for face extraction and the Xception model for classification, produced the best overall performance with an accuracy of 97.4%, precision of 97.0%, and F1-score of 97.2%, demonstrating exceptional robustness in detecting subtle temporal and spatial inconsistencies across frames.

The integrated full-stack application, developed using FastAPI, React, and MongoDB, successfully provided real-time predictions along with confidence scores and stored complete prediction histories for future analysis. Testing with diverse real-world samples showed that the system maintained high reliability across different file formats, compression levels, and recording conditions. Confusion matrices and ROC-AUC analysis indicated strong class separation with minimal false positives and false negatives. The average response time for uploaded media remained within a few seconds, making the platform suitable for practical use in cybersecurity, journalism, law enforcement, and digital forensics. Overall, the results confirm that the proposed multimodal approach significantly enhances detection accuracy and generalization compared with single-modal systems and provides an efficient, scalable, and user-friendly solution for verifying the authenticity of digital media.

VI. CONCLUSION

The proposed research paper, “Multimodal DeepFake Detection Using Deep Learning for Image, Audio, and Video Analysis,” successfully demonstrates the development of an Artificial Intelligence-based system capable of accurately identifying manipulated multimedia content across three major modalities: images, audio, and videos. The system combines Convolutional Neural Networks (CNNs) for image and audio classification with a hybrid MTCNN and Xception pipeline for video analysis, enabling the detection of subtle visual, temporal, and spectral inconsistencies introduced by deepfake generation techniques. Using benchmark datasets such as FaceForensics++, DeepFake Detection Challenge, Celeb-DF, and ASVspoof, the proposed models achieved high detection accuracies of 96.2% for images, 93.7% for audio, and 97.4% for videos, demonstrating the effectiveness of multimodal deep learning for digital media authentication. The integration of FastAPI, React, and MongoDB enabled the deployment of a scalable and user-friendly web application that provides real-time predictions and confidence scores, making the system practical for

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506063|

cybersecurity, journalism, digital forensics, and fraud prevention. Overall, this research confirms that Artificial Intelligence and Deep Learning offer a robust and efficient solution for detecting synthetic media and preserving trust, authenticity, and security in modern digital communication.

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11.
[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
[3] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.
[4] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," in 2019 International Conference on Biometrics (ICB), 2019, pp. 1–8.
[5] E. M. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," IEEE Access, vol. 8, pp. 1–12, 2020.
[6] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5781–5790.
[7] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," arXiv preprint arXiv:2006.07397, 2020.
[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1053–1061.
[9] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 46–52.
[10] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6.
[11] K. K. Raja, N. M. M. Raghavendra, and C. Busch, "Video Presentation Attack Detection in Visible Spectrum Iris Recognition," Pattern Recognition Letters, vol. 82, pp. 15–21, 2016.
[12] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal DeepFake Dataset," in NeurIPS Datasets and Benchmarks Track, 2021.
[13] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.
[14] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning (ICML), 2019.
[15] A. Howard et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.
[16] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[17] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in International Conference on Learning Representations (ICLR), 2021.
[18] X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," in Interspeech, 2020.
[19] 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.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details