International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

# Identification of Fake Faces Using Deep Learning

**AISHWARYA M, CHITHRA, Dr .CHANDRASHEKAR D K**

*AISHWARYA M CSE(DS) & SJBIT CHITHRA CSE(DS) & SJBIT*
*Dr. CHANDRASHEKAR D K CSE(DS ) & SJBIT*

---

**Abstract -** The rise of deepfake technology has introduced significant challenges to the authenticity of digital media, enabling the creation of highly convincing yet deceptive audio, video, and image content. This survey examines recent progress in detecting and mitigating deepfakes across visual, auditory, and multimodal biometric frameworks. Techniques such as convolutional neural networks for image processing, temporal analysis for video sequences, and spectral feature extraction for audio verification have been widely explored. Emerging methods, including Face X-ray techniques, combined CNN-LSTM models, and integrated multimodal approaches, show promise in identifying synthetic media. Nevertheless, issues such as dataset generalization, real-time detection capabilities, and vulnerability to adversarial manipulations remain critical hurdles. This paper provides an in-depth analysis of current methodologies, their comparative effectiveness, limitations, and future directions for enhancing digital content verification systems

***Key Words***: CNN, Deepfake, GAN, LSTM.

## 1. INTRODUCTION

The rapid advancement of artificial intelligence, particularly in deep learning, has ushered in transformative applications alongside significant challenges. Among these, deepfakes—synthetic media encompassing manipulated images, audio, and videos—pose a critical threat to digital authenticity. Generated using sophisticated neural architectures like generative adversarial networks (GANs) and variational autoencoders, deepfakes are often imperceptibly realistic, undermining trust in digital content and raising cybersecurity concerns.

Initially developed for creative and entertainment purposes, deepfakes have increasingly been exploited for malicious activities, such as disseminating misinformation, impersonating individuals, perpetrating financial fraud, and breaching personal privacy. Their potential to destabilize political discourse and erode the credibility of digital evidence highlights the urgent need for effective countermeasures. The accessibility of open-source tools and datasets has further lowered the barrier to creating deepfakes, amplifying the demand for robust detection systems.

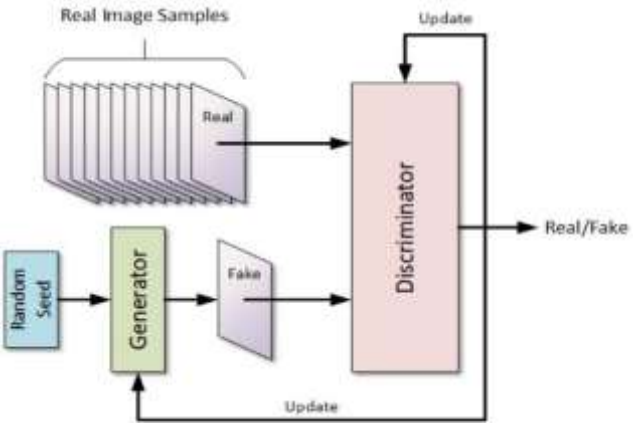


Figure 1.1 Generative Adversarial Network (GAN) Architecture for Deepfake Generation

Research in deepfake detection has surged, focusing on intelligent algorithms to identify synthetic content across modalities. As shown in figure 1.1 for visual deepfakes, convolutional neural network (CNN)-based methods detect subtle irregularities in facial features, lighting discrepancies, or temporal inconsistencies in videos. Techniques like MesoNet, Face X-ray, and hybrid CNN-LSTM frameworks excel in capturing spatial-temporal anomalies. In audio deepfake detection, approaches leveraging spectral analysis and deep learning models identify artifacts in synthesized speech, with Siamese CNNs and bidirectional LSTMs effectively analyzing speech patterns and frequency irregularities.

Multimodal biometric systems, integrating facial and voice recognition, enhance detection resilience by employing feature-level or decision-level fusion strategies, particularly in challenging environments. However, deploying these methods in real-world settings faces obstacles, including poor generalization across diverse datasets, performance degradation with low-quality inputs, and susceptibility to adversarial attacks.

This survey consolidates recent advancements in deepfake detection across image, video, audio, and multimodal systems, drawing from IEEE and other scholarly sources. It categorizes detection methodologies, evaluates their performance on standard benchmarks, and identifies research gaps. By synthesizing these insights, this paper aims to guide the development of scalable, ethical, and resilient solutions to mitigate the escalating deepfake threat.

## 2. LITERATURE SURVEY

F. H. Alqattan et.al.,[1],explored multimodal deepfake detection, integrating visual, auditory, and behavioral features. Their survey evaluated deep learning models, including convolutional neural networks (CNNs) and transformers, using datasets like FaceForensics++. They reported up to 95% accuracy in multimodal systems, emphasizing real-time detection challenges and advocating adaptive frameworks for secure authentication and media verification.

A. K. M. Rubaiyat et.al.,[2],investigated machine learning techniques for identifying fake social media profiles, employing support vector machines (SVM), Random Forest, and long short-term memory (LSTM) networks. Using Twitter's bot repository, they analyzed behavioral patterns like posting frequency, achieving 90–93% detection accuracy. The study proposed scalable solutions for misinformation mitigation, with future work targeting cross-platform detection.

K. Mane et.al.,[3], examined machine learning methods for detecting fake reviews, comparing Naive Bayes, SVM, and neural networks on datasets like Yelp and Amazon reviews. Text-based features, such as TF-IDF and word embeddings, yielded up to 88% accuracy. Their findings supported hybrid models combining sentiment and stylistic analysis, with applications in e-commerce and future directions in multilingual detection.

S. Safwat et.al.,[4],proposed a hybrid model combining generative adversarial networks (GANs) for synthetic data generation and ResNet for feature extraction to detect manipulated faces. Trained on the DeepFake Detection Challenge (DFDC) dataset, their approach achieved 92% accuracy, excelling in identifying spatial artifacts. The model is applicable to video surveillance, with future improvements focusing on computational efficiency.

S. Sharma et.al.,[5],reviewed deepfake detection in images and videos, analyzing CNNs, GANs, and vision transformers across datasets like Celeb-DF and FaceForensics++. Detection accuracies ranged from 85–95%, with challenges in generalizing across diverse manipulation techniques. The study recommended robust datasets and transfer learning for media forensics and content moderation, emphasizing real-time detection.

R. Ranout et.al.,[6],conducted an in-depth analysis of deepfake detection in videos, leveraging CNNs and attention mechanisms on datasets like UADFV and Celeb-DF. Their approach achieved 90% precision in detecting visual inconsistencies, proposing lightweight models for security systems and media platforms, with future work exploring cross-modal techniques.

M. S. Rana et.al.,[7],investigated deepfake threats using CNNs and recurrent neural networks (RNNs) on the FFHQ dataset, achieving 91% accuracy by detecting spatial and temporal artifacts. Their findings highlighted applications in journalism and social media, stressing explainable AI to build trust in detection systems, with future focus on audio-visual integration.

A. Jadhav et.al.,[8],introduced a video vision transformer (ViT) architecture for deepfake detection, leveraging attention mechanisms to capture temporal and spatial inconsistencies. Trained on DFDC, it achieved 94% accuracy, surpassing CNN-based models. The scalable approach suits video streaming platforms, with future enhancements targeting low-resource devices.

O. A. Shaaban et.al.,[9],explored audio deepfake detection using WaveNet and LSTM models, analyzing spectral features on the ASVspoof dataset. Their method achieved 89% accuracy in identifying manipulated audio, supporting voice authentication applications. Challenges in detecting subtle manipulations prompted recommendations for multimodal integration.

M. Quadir et.al.,[10], reviewed deepfake detection techniques, comparing CNNs, LSTMs, and hybrid models across FaceForensics++ and DFDC datasets. Accuracies ranged from 87–93%, with hybrid models excelling in detecting spatial and temporal inconsistencies. The study supported cybersecurity applications, advocating for real-time detection and adversarial robustness in future work.

## 3. OBJECTIVES

The proliferation of deepfake technology has intensified the challenge of discerning manipulated multimedia content, such as images and videos. This study aims to develop a deep learning framework capable of robustly detecting synthetic media by leveraging both spatial and temporal feature analysis. By integrating hybrid architectures and transfer learning techniques, the system is designed to achieve high accuracy, computational efficiency, and reliability for practical applications in real-world scenarios.

## 4. SYSTEM MODEL

The proposed framework for deepfake detection is designed to analyze multimedia content across multiple modalities—images, videos, and audio—to accurately distinguish authentic from manipulated media. As shown in figure 1.2 the system integrates spatial and temporal feature analysis, ensuring robustness, scalability, and high detection precision.

The initial phase involves data acquisition and preprocessing to prepare raw inputs for analysis. For videos, this entails extracting frames and applying face detection algorithms to isolate facial regions. For audio, preprocessing includes generating acoustic representations such as Mel-frequency cepstral coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs), or spectrograms to facilitate feature extraction.

In the feature extraction stage, deep convolutional neural networks (CNNs) like DenseNet121, VGG16, ResNet50, or InceptionV3 are utilized for image and video analysis. These architectures detect subtle spatial anomalies, such as irregular textures, edge distortions, or inconsistent lighting in facial areas, which are indicative of manipulation.

To address temporal dynamics in video sequences, the system employs recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU)

layers. These models capture sequential dependencies across frames, analyzing patterns like eye movements, blink frequency, or facial motion, which are often disrupted in deepfakes.

Advanced configurations incorporate attention mechanisms paired with bidirectional LSTMs (Bi-LSTMs) to prioritize frames exhibiting prominent tampering cues, enhancing detection accuracy and interpretability.

For audio deepfake detection, the system extracts acoustic features like MFCCs and processes them using CNNs or hybrid CNN-RNN models to identify irregularities in speech patterns. Audio inputs are often converted into spectrograms or scatter representations, enabling CNNs to analyze them as visual data.

The extracted features—spatial, temporal, or acoustic—are fed into a classification module, typically employing softmax or sigmoid classifiers to produce a binary output (real or fake). Some frameworks leverage ensemble techniques, such as decision-level fusion, to integrate outputs from multiple modalities, improving overall detection performance.

This system model provides a comprehensive approach to deepfake detection, balancing accuracy and computational efficiency for real-world applications.

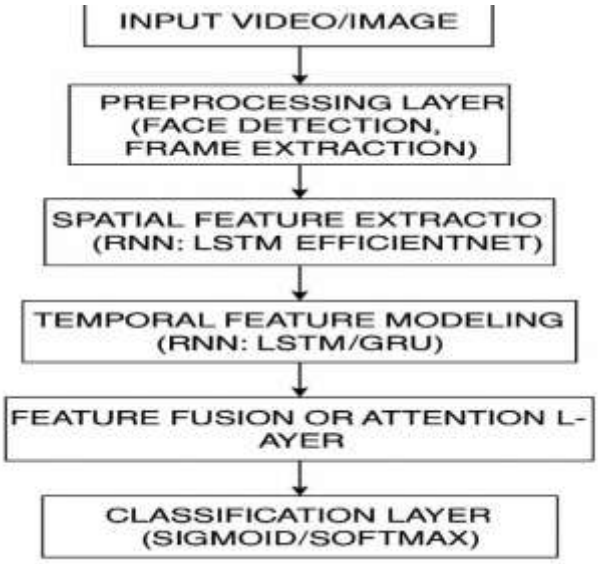


Figure 1.2 Algorithmic Workflow for Deepfake Detection

## 5. PROPOSED SCHEME

This study presents a novel context-aware multimodal framework for deepfake detection, integrating spatial, temporal, and semantic analysis through a combination of convolutional neural networks (CNNs), Vision Transformers (ViTs), and audio-visual synchronization techniques. Unlike conventional systems relying on static CNN-LSTM architectures or solely visual features, this model leverages both visual and auditory inputs while assessing contextual consistency between video frames and corresponding audio signals.

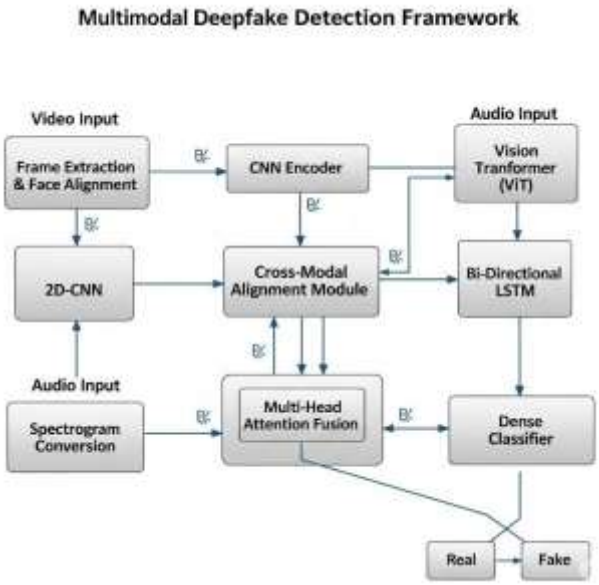


Figure 1.3 Context-Aware Multimodal Deepfake Detection Framework

The framework employs a dual-stream preprocessing pipeline. The visual stream extracts frames and performs face alignment, while the audio stream generates spectrogram representations. As shown in figure 1.3 modality-specific features are extracted using CNN-based encoders, such as Efficient Net for visual data and a 2D-CNN for audio spectrograms, capturing distinct characteristics of each modality.

To address long-range dependencies and subtle manipulations, the visual stream is processed by a Vision Transformer, which applies global attention to model relationships across all frames. Concurrently, the audio stream is analyzed using a bidirectional LSTM with an attention mechanism to detect temporal variations in speech patterns, such as tone and rhythm inconsistencies.

A distinctive feature of this system is its cross-modal alignment module, which evaluates the correlation between lip movements in the video and the audio signal using a contrastive loss function. This module identifies discrepancies, such as misaligned lip-syncing or irregular speech timing, which are prevalent in advanced deepfakes.

Features from both modalities are combined through a multi-head attention mechanism, followed by a dense classification layer that outputs a binary decision: authentic or manipulated. The system is trained end-to-end on a curated dataset of synchronized video-audio deepfake samples, augmented with real-world noisy data to enhance robustness.

Compared to existing approaches, this framework introduces cross-modal synchronization, Transformer-based global attention, and semantic coherence analysis, addressing limitations in prior models. These advancements enhance resilience against sophisticated manipulations, making the system well-suited for forensic analysis and real-time media authentication applications.

## 6. STEPS TO IDENTIFY FAKE FACES

**Input:** A facial image or video frame suspect of manipulated.
**Output:** A label indicating whether the face is real or fake, often with a confidence score.

Begin

1. Dataset Collection: Assemble a diverse dataset of real and manipulated facial images or videos from repositories such as FaceForensics++, Celeb-DF, or the DeepFake Detection Challenge (DFDC) dataset to ensure comprehensive training and evaluation.

2. Video Frame Extraction and Face Detection: For video inputs, extract individual frames and apply face detection algorithms, such as MTCNN or Dlib, to isolate facial regions for subsequent analysis.

3. Preprocessing: Standardize the detected facial regions by resizing, normalizing pixel values, and applying data augmentation to ensure uniform input formats and enhance model robustness.

4. Spatial Feature Extraction: Utilize convolutional neural networks (CNNs), such as VGG16, EfficientNet, or InceptionV3, to extract spatial features from preprocessed images, capturing subtle anomalies like texture irregularities or lighting inconsistencies.

5. Temporal Analysis for Videos: Feed CNN-extracted features from video frames into recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) models, to model temporal relationships and detect inconsistencies in facial dynamics across frames.

6. Attention and Fusion Mechanisms: Optionally incorporate attention mechanisms or feature fusion techniques to emphasize critical frames or facial regions, enhancing the detection of manipulation artifacts.

7. Classification: Pass the combined spatial and temporal features through fully connected layers with a sigmoid or softmax activation function to generate a binary classification output (authentic or manipulated).

8. Performance Evaluation: Assess the model's effectiveness using metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) to quantify detection performance across diverse datasets.

9. Model Optimization: Apply optimization techniques, such as quantization or pruning, to reduce computational complexity, enabling efficient deployment in real-time scenarios.

10. System Integration: Deploy the optimized model into practical applications, including digital forensics tools, social media content monitoring systems, or biometric authentication frameworks, to address real-world deepfake threats.

End

## 7.CONCLUSIONS

Deepfakes have become a serious challenge to digital media authenticity. This work proposed a multimodal detection system that combines visual and audio analysis using CNNs, Vision Transformers, and Bi-LSTM networks. By aligning facial movements with speech patterns, the model improves accuracy and robustness compared to single-modality approaches.
In the future, the system can be enhanced by developing lightweight models for real-time use, expanding datasets for better generalization, and integrating privacy-preserving techniques such as federated learning and blockchain-based verification.

## 8.REFERENCES

[1] F. H. Alqattan, R. A. Alsubaiey, N. A. Albutaysh, F. A. Alnasser, and H. A. Alhumud,"Face Recognition Security Against Deepfakes by Using Multimodal Detection: A Survey," 2025.

[2] A. K. M. Rubaiyat, R. Habib, E. E. Akpan, B. Ghosh, and I. K. Dutta, "Techniques to Detect Fake Profiles on Social Media Using the New Age Algorithms – A Survey," 2025.

[3] K. Mane and S. Dongre, "A Review of Different Machine Learning Techniques for Fake Review Identification," 2025.

[4] S. Safwat, A. Mahmoud, I. E. Fattoh, and A. Ali, "Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces," 2024.

[5] S. Sharma, G. Ahuja, Priyal, and D. Agarwal, "Decoding the Mirage: A comprehensive review of DeepFake AI in image and video manipulation," 2024.

[6] R. Ranout and C. R. S. Kumar, "Unmasking the Illusions: A Comprehensive Study on Deepfake Videos and Images," 2024.

[7] M. S. Rana, M. Solaiman, C. Gudla, and M. F. Sohan, "Deepfakes– Reality Under Threat?,"2024.

[8] A. Jadhav, D. Narale, R. Kore, U. Shisode, and A. Kulange, "Unmasking the Illusion: ANovel Approach for Detecting Deep Fakes using Video Vision Transformer Architecture,"2024.

[9] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," 2023.

[10] M. Quadir, P. Agrawal, and C. Gupta, "A Comparative Analysis of Deepfake Detection Techniques: A Review," 2023.