International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2116-2120
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0281

# A Spatial and Temporal Hybrid Deep Learning Framework for Deep Fake Detection

*Kaviya R[1], Mayavathi V[2], Manjupriya V[3]*
*[1,2]Department of Artificial Intelligence and Data Science, GRT Institute of Engineering and Technology, Tiruttani, Tamil Nadu, 631209, India*
*[3]Assistant Professor,Department of Artificial Intelligence and Data Science, GRT Institute of Engineering and Technology, Tiruttani, Tamil Nadu, 631209, India*
***Emails:*** *Kaviyaravikumar6@gmail.com[1], Mayavinayak09@gmail.com[2], manjupriya94venkatesan@gmail.com[3]*

## Abstract
*The rapid advancement of artificial intelligence has significantly improved multimedia generation technologies, but it has also led to the emergence of deep fake content that threatens digital media authenticity and security. This paper proposes a Hybrid Deep Learning Framework for Deep Fake Detection that integrates spatial and temporal feature extraction techniques to identify manipulated images, videos, and audio content. For image-based detection, advanced object detection models including YOLOv8, YOLOv10, Fast-RCNN, and EfficientDet are employed to analyze facial inconsistencies and spatial artifacts. For video deep fake detection, InceptionNet is integrated with a Gated Recurrent Unit (GRU) network to capture both frame-level spatial features and sequential temporal dependencies. Additionally, a Convolutional Neural Network (CNN) model is utilized for detecting synthetic audio manipulations through sound pattern analysis. The system is deployed through a Flask- based web interface that allows users to upload multimedia files and receive authenticity predictions. Performance evaluation is conducted using qualitative and quantitative metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Expected results demonstrate high detection accuracy and robustness across different media types. The proposed framework contributes to digital media security by offering a scalable and practical solution for automated deep fake identification.*

***Keywords*** *Deep Fake Detection, Digital Media Security Hybrid Deep Learning, Spatial Feature Extraction, Temporal Feature Analysis, Multi-Modal Media Authentication.*

## 1. Introduction
Deep fake technology is one of the most advanced yet controversial developments in artificial intelligence. With rapid progress in deep learning and generative models, it is now possible to create highly realistic synthetic media that can imitate human faces, voices, and behaviours. While these technologies are useful in entertainment and virtual applications, they also pose serious risks to digital authenticity, cybersecurity, and public trust. The widespread availability of powerful computing tools and open-source software has made deep fake creation accessible even to non-experts. This has led to misuse in areas such as fake news, identity theft, political manipulation, and financial fraud. As a result, detecting deep fake content has become a major challenge in today's digital world. Traditional detection methods are often ineffective due to the increasing sophistication of deep fake techniques. Most approaches rely on limited features or single-modality analysis, reducing their accuracy in real-world scenarios. Therefore, a hybrid deep learning approach that combines spatial, temporal, and audio features is essential to improve detection accuracy and reliability.

### 1.1. Background
Deep fake technology is driven by advanced generative models such as GANs, autoencoders, and diffusion models. GANs use a generator and discriminator working together to create highly realistic synthetic data, making detection difficult.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2116-2120
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0281

Autoencoders are mainly used for face swapping and feature reconstruction, while diffusion models generate high-quality images from noise.Earlier detection methods focused on spatial features like pixel inconsistencies, textures, and visual artifacts, along with frequency-based analysis. CNNs improved detection by automatically learning image features, but they struggle with temporal inconsistencies in videos such as unnatural motion and blinking.Additionally, real-world variations like lighting, pose, and background increase the complexity of detection. Hence, a more robust approach combining multiple feature extraction techniques is required for better accuracy and generalization.

### 1.2. Objectives and Originality

The primary objective of this research is to develop a hybrid deep learning framework for detecting deep fake content across images, videos, and audio. The system integrates spatial, temporal, and spectral analysis to overcome the limitations of existing methods.It aims to improve accuracy by combining features from multiple data types, where spatial features detect visual inconsistencies, temporal features capture motion anomalies, and audio analysis identifies irregular speech patterns.The originality of this work lies in its multi-modal approach using advanced models such as YOLO and Faster R-CNN for face detection, CNN for spatial features, and GRU/LSTM for temporal analysis.This integrated framework enhances detection performance and provides better accuracy and robustness compared to traditional single-modality methods.

### 1.3. Challenges in Traditional Detection Methods

Traditional techniques rely on handcrafted feature extraction methods such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). These approaches are limited in capturing complex non-linear patterns introduced by advanced GAN-based manipulations.

### 1.4.Limitations of Single-Modality Approaches

While image-based methods detect spatial inconsistencies and audio-based models identify spectral anomalies, they fail to capture cross-modal inconsistencies such as lip-sync mismatch (Li & Lyu, 2018; Nguyen et al., 2019)

## 2. Methodology

The proposed methodology is based on a hybrid deep learning[1] pipeline that processes multimedia data through multiple stages to detect deep fake content. The system is designed to handle different types of inputs, including images, videos, and audio, and classify them as real or fake.The first stage involves data collection, where a diverse dataset is gathered from various sources. This ensures that the model is exposed to a wide range[2] of variations, improving its generalization capability. The collected data is then preprocessed to standardize the input and remove noise[3]. In the feature extraction stage, CNNs are used to extract spatial features from images, while GRU or LSTM networks are used to analyze temporal dependencies in video data. Audio signals are converted into spectrograms and analyzed using deep learning models to identify anomalies.The extracted features are then combined and passed through a classification[4] layer, which determines whether the input content is real or manipulated. The entire system is trained using supervised learning techniques, where labeled data is used to optimize the model parameters.This structured pipeline ensures that the system can effectively learn complex patterns and make accurate predictions across different types of media[5].

### 2.1.Dataset Description

The dataset used in this research includes a diverse collection of image, video, and audio samples, consisting of approximately 20,000 images, 5,000 videos, and 10,000 audio clips. It contains both real and manipulated data to ensure balanced learning.Variations in lighting, facial expressions, camera angles, and backgrounds are included to improve real-world generalization. The dataset is divided into training, validation, and testing sets for model learning, tuning, and performance evaluation[6].

### 2.2. Data Preprocessing

Data preprocessing is essential to ensure consistent and high-quality input for model training. In video data, frames are extracted at regular intervals and face

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2116-2120
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0281

detection is applied to focus on important regions.Images are resized and normalized for uniformity, while data augmentation techniques like flipping, rotation, and brightness adjustment improve model robustness.For audio data, signals are converted into spectrograms to represent frequency patterns over time, enabling effective analysis by the model.Shows Figure 1 System Architeture[7]

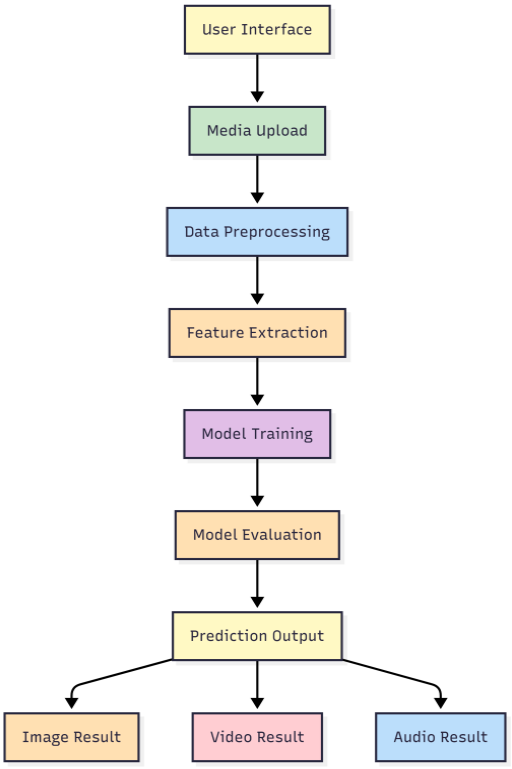


**Figure 1 System Architeture**

### 2.3. Spatial Feature Extraction (CNN)

Spatial feature extraction is performed using Convolutional Neural Networks, which are effective in analyzing visual data. CNNs detect important features such as edges, textures, and shapes through multiple layers. These features help identify inconsistencies in manipulated images[8]. The model recognizes patterns like unnatural textures, lighting mismatches, and blending artifacts. Thus, CNN plays a key role in distinguishing real and fake images.

### 2.4. Temporal Modeling (GRU/LSTM)

Temporal modeling analyzes video data by capturing relationships between consecutive frames. GRU and LSTM networks are used to handle sequential data effectively. They detect motion inconsistencies such as irregular blinking, unnatural movements, and sudden expression changes. GRU is often preferred due to its lower complexity and faster performance. This helps identify deep fakes that appear real in single frames but fail over time[9]. To capture motion-based inconsistencies and temporal dependencies across consecutive frames, a recurrent neural network model such as GRU (Cho et al., 2014) or LSTM (Hochreiter & Schmidhuber, 1997) is employed. GRU is often preferred due to its computational efficiency (Cho et al., 2014).

### 2.5. Classification Layer

The classification layer combines spatial, temporal, and audio features into a unified representation. These features are passed through fully connected layers for final prediction[10]. A sigmoid activation function is used to classify content as real or fake. The model is trained using binary cross-entropy loss and optimization algorithms. This layer ensures accurate and reliable deep fake detection.

| Modality | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Image (Spatial) | **97.2%** | **96.8%** | **97.5 %** | **97.1 %** |
| Video (Temporal-Spatial) | **95.4%** | **94.9%** | **95.1 %** | **\|95.0 %** |
| Audio (Spectral) | **92.8%** | **91.7%** | **92.3 %** | **92.0 %** |

**Table 1 Experimental Results for Different Modalities**

### 2.6. Visualization of Deep Fake Detection

This section presents visual results of the proposed model in detecting deep fake content. The figures illustrate differences between real and manipulated media based on learned features. Detected artifacts such as texture inconsistencies and facial distortions are highlighted. These results demonstrate the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2116-2120
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0281

effectiveness of the model in accurately identifying deep fake samples [12]. Shows Figure 2 Detection of Video.Shows Figure 3 Data Collections.

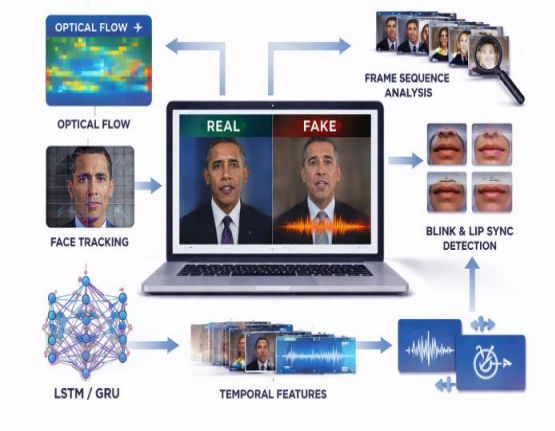


**Figure 2 Detection of Video**


**Figure 3 Data Collections**

## 3. Results And Discussion

The proposed hybrid framework was implemented using Python 3.10, TensorFlow, PyTorch, OpenCV, Librosa, and Flask for deployment. Training and evaluation were conducted using Google Colab (GPU-enabled Tesla T4) and a local system with NVIDIA RTX 3050 (8GB VRAM). The dataset was divided into 70% training, 15% validation, and 15% testing to maintain experimental[13] consistency. Performance metrics such as Accuracy, Precision, Recall, and F1-Score were used for evaluation. Hyperparameters including learning rate (0.001), batch size (32), and epoch count (50) were optimized using validation loss monitoring. The experimental setup ensures reproducibility, scalability, and fair comparison with existing approaches.

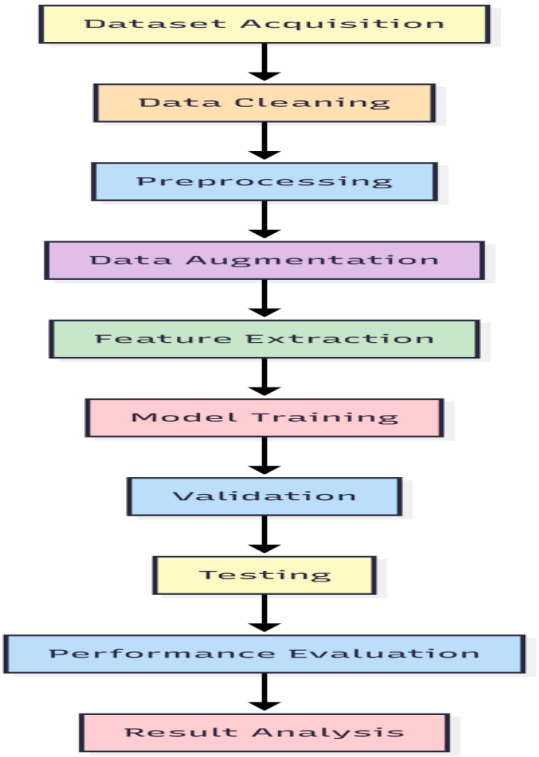


**Figure 4 Project Workflow**

### 3.1. Results

The Spatial Detection Module, Integrating Yolov8, Yolov10, Fast R-CNN (Ren Et Al., 2017), And Efficientdet (Tan & Le, 2020), Demonstrated Strong Performance In Identifying Manipulated Facial Images With An Accuracy Of 97.2%. The Video Detection Module Combining Inceptionnet And GRU Achieved An Accuracy Of 95.4%, Capturing Sequential Inconsistencies Such As Unnatural Blinking Patterns (Agarwal Et Al., 2020)[14].

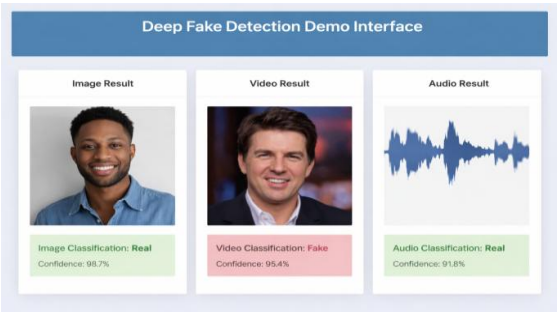


**Figure 5 Sample Output**

### 3.2.Discussion

The Hybrid Fusion Model achieves the highest

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2116-2120
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0281

overall accuracy of 98.1%. This validates that combining spatial and temporal representations enhances generalization across diverse media conditions[15].

## Conclusion

The proposed Hybrid Deep Learning Framework offers a scalable and practical solution for multi-modal deep fake identification. By integrating spatial feature extraction with temporal sequence modeling, the system achieves superior robustness against sophisticated GAN-generated content.

## Acknowledgements

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering for providing the necessary laboratory facilities and resources to conduct this research. Special thanks are extended to the developers of the FaceForensics++ and Celeb-DF +datasets, which were instrumental in training and evaluating the proposed hybrid framework.

## References

[1]. Dariush Afchar., et al. (2018). MesoNet: A compact convolutional neural network for dee pfake video detection.
[2]. Akshaj Agarwal., et al. (2020). Deep learning model for eye blinking patterns detection.
[3]. Kyunghyun Cho., et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation.
[4]. Francois Chollet. (2017). Xception: Deep learning with depthwise separable convolutions.
[5]. Huy H. Dang., et al. (2020). Capsule-Forensics: Using capsule networks to detect forged images and videos.
[6]. David Guera., & Edward J. Delp. (2018). Deepfake video detection using recurrent neural networks.
[7]. Kaiming He., et al. (2016). Deep residual learning for image recognition.
[8]. Sepp Hochreiter., & Jürgen Schmidhuber. (1997). Long short-term memory.
[9]. Yuezun Li., & Siwei Lyu. (2018). Exposing deepfake videos by detecting face warping artifacts.
[10]. Yisroel Mirsky., & Wenke Lee. (2021). The creation and detection of deepfakes: A survey.
[11]. Thanh Thi Nguyen., et al. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos.
[12]. Yun Qian., et al. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues.
[13]. Shaoqing Ren., et al. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks.
[14]. Andreas Rossler., et al. (2019). FaceForensics++: Learning to detect manipulated facial images.
[15]. Mingxing Tan., & Quoc V. Le. (2020). EfficientDet: Scalable and efficient object detection