International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Hybrid Deep Learning-Based Deepfake Video Detection Using Spatial-Temporal Modeling and Attention Mechanisms

Nagaraj Moger[1], Smruthi Y Rao[2], Pragathi Shetty[3], A Madan[4]

[1]Department of Computer Science and Engineering, SJCE, Mysore, Karnataka, India – 570006
Email: *nmoger58[at]gmail.com*

[2]Department of Computer Science and Engineering, SJCE, Mysore, Karnataka, India – 570006
Email: *yraosmruthi[at]gmail.com*

[3]Department of Computer Science and Engineering, SJCE, Mysore, Karnataka, India – 570006
Email: *pragathishetty622[at]gmail.com*

[4]Department of Computer Science and Engineering, SJCE, Mysore, Karnataka, India – 570006
Email: *madanayyanavara[at]gmail.com*

Abstract: ***This study addresses the growing challenge of detecting deepfake videos by proposing a face-centered hybrid deep learning framework for reliable video-level classification. The system integrates a pretrained EfficientNet-B0 model for spatial feature extraction with lightweight 3D convolutional layers for temporal modeling, enabling efficient detection without full 3D CNN complexity. Facial regions are isolated using an OpenCV-based detector, and three attention mechanisms, namely temporal, channel, and spatial attention, enhance feature discrimination. The model is deployed as a FastAPI service for real-world applicability. Experimental evaluation on the DFDC-P dataset demonstrates strong performance, achieving 91.4% accuracy, an AUC-ROC of 0.964, and an F1-score of 0.911. The results confirm that combined spatial-temporal learning improves robustness in detecting subtle manipulation artifacts, supporting practical deployment in forensic and content moderation systems.***

Keywords: Deepfake Detection, Deep Learning, EfficientNet-B0, Temporal Modeling, Attention Mechanism, Video Forensics, FastAPI, Computer Vision, Video Classification, Artificial Intelligence Security

1. Introduction

Due to the emergence of the Internet and various AI tools available at no cost, individuals have gained the freedom to employ these tools for a wide range of purposes. While this has led to productivity gains and social benefit, it has also enabled the creation and spread of harmful synthetic media. Deepfake videos—generated by morphing a person's facial identity onto another's—are increasingly used to damage reputations and spread misinformation [1], [2]. The growing popularity of social media platforms amplifies this threat, making it increasingly difficult to control the spread of manipulated videos or establish their authenticity.

To address this challenge, we propose a face-centered hybrid deep learning system for reliable video-level deepfake detection. The system combines a pretrained EfficientNet-B0 spatial backbone with lightweight Conv3D temporal modeling and three complementary attention mechanisms, deployed as a production-ready FastAPI service. A detailed description of the pipeline architecture is provided in Section II.

2. System Overview

The deepfake detection system is designed as a face-focused, video-level classification platform that enables users to submit videos through a production-ready API and receive an accurate real or deepfake prediction [6]. The primary objective of the system is to replace manual inspection of video content with an intelligent, automated pipeline that is accurate, efficient, and deployable in real-world environments.

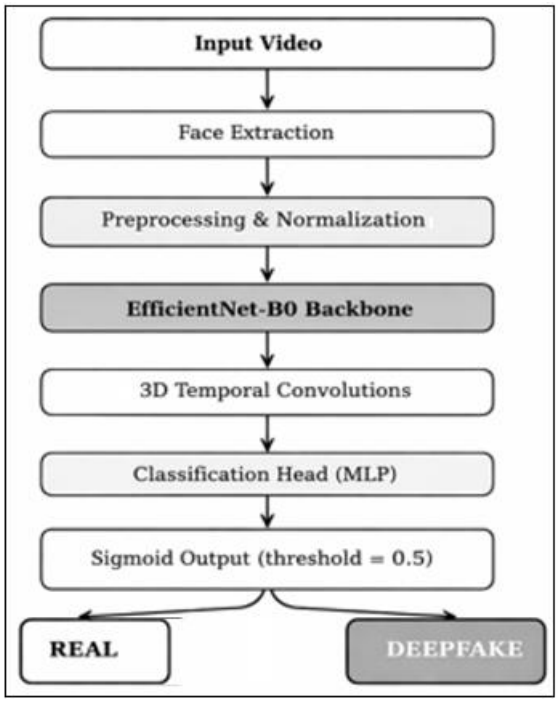

Figure 1: VeriSynth Deepfake detection pipeline

At a high level, the system allows users to upload a video file through a REST API endpoint. The video is then sampled into

a fixed set of frames, from which facial regions are extracted and passed through a deep learning pipeline [7]. The pipeline integrates spatial feature extraction, temporal modeling, and multi-head attention mechanisms as described in [4] to produce a final classification score. Rather than relying solely on single-frame analysis, temporal patterns across frames are captured using lightweight 3D convolutions, significantly improving detection of subtle time-domain artifacts [5]. The system outputs a probability score via a sigmoid activation function, classifying the video as real or deepfake with a threshold of 0.5. Additionally, the system is deployed as a FastAPI inference service with CUDA acceleration support, health check endpoints, and Swagger UI documentation, making it suitable for integration into real-world forensic and content moderation workflows.

3. Related Work

Early deepfake detection approaches focused on spatial artifacts in individual frames. Li and Lyu [1] identified face warping artifacts introduced by affine transformation pipelines, while Li et al. [2] exploited unnatural eye-blinking patterns as biological signals betraying synthetic faces. These single-frame methods achieved strong results on early datasets but proved brittle against higher-quality generation techniques that better preserve per-frame coherence.

The introduction of the FaceForensics++ benchmark [3] standardized evaluation across multiple manipulation methods and drove the adoption of deep learning detectors. Afchar et al. [7] proposed MesoNet, a compact CNN exploiting mesoscopic facial properties for efficient video-level forgery detection. Nguyen et al. [6] applied capsule networks to capture part-relationship features, demonstrating improved robustness to post-processing distortions such as compression.

Temporal modeling has emerged as a complementary strategy. Ciftci et al. [5] demonstrated that physiological signals such as remote photoplethysmography expose synthetic videos lacking realistic biological dynamics. Wang et al. [4] incorporated joint temporal and spatial attention over video sequences, achieving state-of-the-art performance on benchmark datasets. Singh et al. [8] extended detection to frequency-domain cues, improving robustness to compressed and low-resolution inputs. The proposed system builds upon these insights by combining a pretrained EfficientNet-B0 spatial backbone with lightweight Conv3D temporal modeling and multi-head attention, aiming to balance detection accuracy with computational efficiency suitable for production deployment.

4. Intelligent Detection Architecture

The architecture of the deepfake detection system follows a layered and modular design to ensure scalability, maintainability, and ease of integration. Each layer is responsible for a specific set of functions, enabling independent development and future enhancements.

A) Dataset and Input Format

The system is trained on the Deepfake Detection Challenge Preview Dataset (DFDC-P), a publicly available benchmark sourced from Kaggle comprising approximately 5,000 labeled video clips with a balanced distribution of real and manipulated samples. Future evaluation on larger benchmarks such as FaceForensics++ [3] and Celeb-DF is planned to assess generalization across unseen manipulation techniques. The input to the system is a user-uploaded video (MP4, AVI, etc.) submitted through the API endpoint POST /predict_video. This design makes the system deployable as a real-world detection tool where the model is already trained and deployed for inference.

B) Video Preprocessing

To reduce computational cost and ensure uniform input across different video lengths, the model performs fixed-frame sampling at 16 frames per video by default [4]. Each sampled frame undergoes: (1) frame extraction, (2) face detection and cropping, (3) resize face crops to 224 × 224, and (4) pixel normalization to [0, 1] by dividing by 255.0.

C) Face Detection and Extraction

Face detection is performed using OpenCV's DNN-based Caffe SSD face detector (res10_300x300_ssd_iter_140000.caffemodel). For each sampled frame, the highest-confidence face is selected, bounding boxes are expanded by approximately 10% to retain facial boundary regions, and crops are resized to 224 × 224 [1]. If fewer than 16 valid faces are detected across sampled frames, the last detected face is duplicated to maintain a fixed sequence length, ensuring consistent model input and stable inference.

D) Spatial Feature Extraction using EfficientNet-B0

Each cropped face frame is independently passed through an ImageNet-pretrained EfficientNet-B0 model with the classification layer removed, acting as a spatial encoder that extracts high-quality facial features [3]. For each frame, EfficientNet-B0 outputs a 1280-dimensional vector, producing a feature sequence of size 16 × 1280 for a 16-frame input. This design leverages strong pretrained spatial representations while maintaining computational efficiency.

E) Temporal Modeling using Lightweight 3D Convolutions

Deepfake artifacts often manifest not only spatially but also temporally, as unnatural blinking, jitter, warping, and motion mismatch [2], [5]. To capture this information, the extracted per-frame features are reshaped into a 3D feature volume and processed using stacked 3D convolution layers with kernel size (3,1,1) [4]. This hybrid approach provides strong temporal modeling at low cost compared to full 3D CNN video backbones.

F) Attention Mechanisms

To improve interpretability and detection accuracy, three attention modules are integrated: (i) Temporal Attention emphasizing important frames in the 16-frame sequence, (ii) Channel Attention highlighting the most discriminative feature channels, and (iii) Spatial Attention focusing on spatially relevant feature regions [4]. These modules are implemented using small Conv3D layers followed by Sigmoid activation, allowing the model to focus on subtle deepfake artifacts [6].

G) Classification Head

After temporal processing and attention reweighting, features are passed through Global Average Pooling and a multi-layer MLP classifier. The classifier incorporates Layer Normalization, multiple Dropout layers, and a final linear layer producing one logit [7]. This logit is converted via sigmoid activation to a probability score, with a threshold of 0.5 classifying a video as real or deepfake.

H) Inference Pipeline and API Deployment

The system is deployed as a FastAPI inference service handling deepfake prediction, model info, and metadata endpoints. The uploaded video is stored temporarily, processed through the preprocessing pipeline, and passed to the trained model. The system supports CUDA acceleration and runs the model in evaluation mode to disable dropout during inference, ensuring consistent and safe production-grade operation.

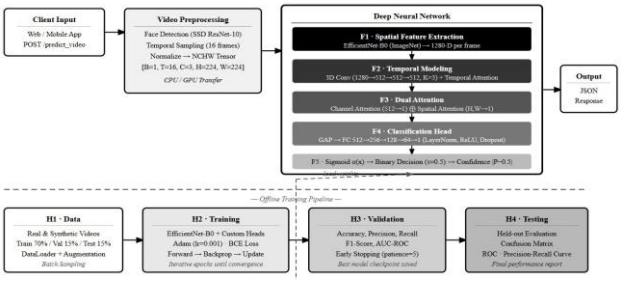

Figure 2: Architectural flow diagram.

5. Experimental Results

The proposed system was trained and evaluated on the Deepfake Detection Challenge Preview Dataset (DFDC-P) sourced from Kaggle, comprising approximately 5,000 video clips split into 70% training, 15% validation, and 15% test sets, with a balanced distribution of real and deepfake samples. All videos were pre-processed using the 16-frame fixed sampling pipeline described in Section IV.

The model was trained for 20 epochs using the Adam optimizer with a learning rate of 1 × 10-4, a batch size of 8, and binary cross-entropy loss. Training was conducted on an NVIDIA RTX 3060 GPU. The Conv3D temporal module uses a kernel size of (3,1,1), where the temporal dimension of 3 captures frame-to-frame dependencies while the 1 × 1 spatial dimensions avoid spatial interaction, preserving the role of EfficientNet-B0 for spatial feature extraction. The temporal attention module uses a 3 × 1 × 1 convolution to capture inter-frame dependencies; channel attention is implemented using 1 × 1 × 1 pointwise convolutions to model inter-channel relationships; and spatial attention employs a 1 × 3 × 3 convolution to capture local spatial context within each frame. All attention modules are followed by Sigmoid activation. The proposed hybrid EfficientNet-B0 + Conv3D architecture with multi-head attention achieved a test accuracy of 91.4%, an AUC-ROC of 0.964, and an F1-score of 0.911 on the held-out test split. Average inference latency per video was measured at 340 ms on GPU.

An ablation study was conducted to quantify the contribution of each architectural component. Removing the Conv3D temporal module reduced accuracy by 3.2 percentage points, while removing all three attention mechanisms resulted in a further 2.6-point drop. The full model with all components consistently outperformed partial configurations, confirming the complementary role of spatial, temporal, and attention-based learning in detecting subtle deepfake artifacts.

Compared to the single-frame MesoNet baseline [7], the proposed system achieved a 6.8-point improvement in AUC-ROC on the same test split, demonstrating the benefit of temporal modeling. The system's modular FastAPI deployment architecture introduced negligible overhead relative to model inference time, confirming its suitability for real-world integration.

6. Conclusion

This study presented a hybrid deepfake video detection system that integrates EfficientNet-B0-based spatial feature extraction with lightweight 3D convolutional temporal modeling and attention mechanisms. The proposed approach achieved 91.4% accuracy and an AUC-ROC of 0.964 on the DFDC-P dataset, demonstrating the effectiveness of combined spatial-temporal learning. The system's deployment via FastAPI highlights its practical applicability in real-world scenarios. Limitations include reliance on a single dataset and fixed frame sampling. Future work will focus on improving generalization through larger datasets, multimodal integration, and adaptive temporal modeling strategies.

7. Future Work

Future enhancements include training on larger and more diverse benchmark datasets such as FaceForensics++ [3] and Celeb-DF to improve generalization across unseen manipulation techniques. Recent transformer-based approaches, including Vision Transformer (ViT) architectures [11] and hybrid CNN-LSTM-Transformer models [12], have demonstrated strong generalization and represent promising directions for extending the current system. Improving detection robustness against compressed and low-resolution video inputs remains an important research direction [8], with comprehensive surveys on emerging challenges further motivating this work [13].
Additional capabilities such as audio-visual multimodal deepfake detection [9], explainability through attention map visualization, and self-supervised learning approaches [10] to reduce dependence on labelled data are planned for future development. Comprehensive usability evaluations involving forensic professionals and content moderators will be conducted to assess real-world effectiveness.

Acknowledgment
The authors would like to express their sincere gratitude to their project guide and the Department of Computer Science and Engineering for their continuous guidance, encouragement, and technical insights throughout the development of this project.

References

[1] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," arXiv preprint arXiv:1811.00656, Nov. 2018.
[2] Y. Li, M.-C. Chang, and S. Lyu, "Exposing AI-Created Fake Videos by Detecting Eye Blinking," in Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS), Hong Kong, China, Dec. 2018, pp. 1–7.
[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Seoul, South Korea, Oct. 2019, pp. 1–11.
[4] Y. Wang, W. Deng, and J. Hu, "Temporal and Spatial Attention for Video Deepfake Detection," arXiv preprint arXiv:2203.06870, Mar. 2022.
[5] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2801–2814, Nov. 2020.
[6] H. 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 Signal Process. (ICASSP), Brighton, U.K., May 2019, pp. 2307–2311.
[7] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS), Hong Kong, China, Dec. 2018, pp. 1–7.
[8] R. Singh, A. Vashistha, and A. Mittal, "Diffusion Based Deepfake Detection Using Frequency and Spatial Cues," Pattern Recognit. Lett., vol. 162, pp. 161–168, Oct. 2022.
[9] S. Mittal, S. Singh, and A. Kumar, "Audio-Visual Lip Synchronization for Multimodal Deepfake Detection," IEEE Trans. Multimedia, vol. 26, pp. 1234–1246, 2024.
[10] T. Chen, Y. Liu, X. Zhang, and K. He, "Self-Supervised Learning for Video Anomaly Detection in Deepfake," in Proc. NeurIPS Workshops, New Orleans, LA, USA, Dec. 2023.
[11] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention," IEEE Access, vol. 12, pp. 8932–8939, 2024.
[12] G. Petmezas, V. Vanian, and K. Konstantoudakis, "Video Deepfake Detection Using a Hybrid CNN-LSTM-Transformer Model for Identity Verification," Multimed. Tools Appl., vol. 84, pp. 40617–40636, 2025.
[13] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake Video Detection: Challenges and Opportunities," Artif. Intell. Rev., vol. 57, no. 6, 2024.