International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

Deep Deception Detector: Exposing AI Generated Fake Video

Shahin Sayyad[1], Maitree Wasnik[2], Anjali Abhang[3], Apurva Chavan[4], Hema Jadhav[5]
[1,2,3,4,5]Artificial Intelligence and Data Science, Vpkbiet, Baramati, India.
Emails: *shahinsayyad1302@gmail.com[1], maitreewasnik2004@gmail.com[2], anjalivabhang@gmail.com[3], apurvachavan306@gmail.com[4], hemakjadhav@gmail.com[5]*

Abstract
Deepfakes are artificially generated videos or images created to falsely portray someone as saying or doing things they never actually did. These manipulated media forms can lead to serious issues, especially when circulated on social media platforms. As a result, detecting deepfakes has become increasingly critical. This project builds on an existing detection method called FAMM, which targets identifying deepfakes, particularly in videos that have been compressed. The original FAMM approach analyzes facial movements by calculating distances and angles between key facial landmarks, such as the eyes, nose, and mouth, and observing how these points change over time. In this earlier method, GRU and SVM models were used to capture the temporal and static changes, and their outputs were combined to determine whether the video was authentic or fake. In our updated approach, we introduce more advanced techniques. We replace the GRU with a Transformer model, which offers improved capabilities in capturing time-based changes in facial movements. Additionally, we implement EfficientNet to extract more precise features from the face images. The data from both models are processed and then combined through a fusion strategy to reach a final classification of whether the video is real or fake. With these advancements, our system demonstrates improved accuracy in detecting deepfakes, even in low-quality or compressed videos. This project highlights how cutting-edge deep learning techniques can better address the spread of deepfakes on social media platforms.
Keywords: *Deepfakes, detection, Transformer, EfficientNet, facial landmarks, feature extraction, fusion.*

1. Introduction
Deepfakes are fake videos or images made to look like someone is saying or doing things they never actually did. These manipulated videos can cause serious problems, especially when shared on social media, because they spread false information and can damage trust. Detecting deepfakes is becoming more important as they become more common. (Figure 1)

Figure 1 DeepFake Image

This project builds on a method called FAMM, which helps find deepfakes, especially in videos that have been compressed. FAMM looks at facial movements by measuring how key points on the face, like the eyes, nose, and mouth, move and change over time. In the original method, GRU and SVM models were used to study these changes and decide if the video was real or fake. In this improved approach, newer techniques are used. The GRU model is replaced with a Transformer model, which is better at analyzing changes in facial movements over time. EfficientNet is also used to get more accurate features from the face. The results from both models are combined to decide whether the video is real or fake. With these improvements, the system is better at detecting deepfakes, even in low-quality or compressed videos. This project shows how advanced deep learning techniques can help address the spread of deepfakes on social media.

2. Literature Survey
The field of deepfake detection has seen significant advances, with multiple approaches exploring spatial,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

temporal, and feature-based inconsistencies. One promising approach analyzes facial muscle motion, combining Gated Recurrent Units (GRU), Support Vector Machines (SVM), and DempsterShafer theory, which shows robustness to compression using a dual classifier and fusion strategy. However, this method faces computational challenges and is limited in addressing temporal discontinuities. Another method leverages an integration of the Convolutional Block Attention Module (CBAM), Video Vision Transformer (VVT), and Depthwise Separable Convolution (DSC) to enhance accuracy and feature extraction efficiency. While this approach improves detection accuracy, it can overfit, complicating deployment on diverse datasets. A distinct approach uses recurrent convolutional structures and Mel-Frequency Cepstral Coefficients (MFCC), alongside Power Spectrograms and Convolutional Neural Networks (CNN), to detect deepfakes across both audio and visual media. Despite its adaptability, high computational demands limit its suitability in low-resource environments, and generalizability to real-world data remains a challenge. Additionally, combining methods like MesoNet and Xception, Discrete Cosine Transform, and MRE-Net effectively captures dynamic spatial-temporal inconsistencies in largescale fake detection. However, the method's complexity, limited generalization to high-quality deepfakes, and potential overfitting on small datasets create limitations in certain scenarios. A dual-stream CNN model, integrating ResNet18, pruning, and Coviar, has shown effectiveness in compressed video detection for social networks. However, the dependency on temporal consistency and the trade-offs introduced by pruning may impact robustness, limiting performance under varying conditions. To capture spatial and temporal inconsistencies across different playback speeds, another technique utilizes Bipartite Group Sampling, Momentary Inconsistency Excitation, and LIE ResNet. While effective, high computational demands and challenges in low-quality video scenarios limit real-world applicability. An unsupervised framework using Photo Response NonUniformity (PRNU), Binary Cross-Entropy, and a SelfSubtract Mechanism provides robustness to manipulation, showing promise for interpretability. However, this method struggles with low-quality video data and high computational costs, which hinder real-time use. An innovative approach incorporating a Complementary Cross Dynamics Fusion Module (CCOFM) excels in detecting both subtle and significant frame changes, but sensitivity to compression and high computational needs pose scalability challenges. Another approach uses the TimeSformer architecture with CNNs and a Global-Local Transformer to capture diverse spatiotemporal views, demonstrating high detection accuracy but requiring significant processing power, which limits its real-time applications. Lastly, a Self-Subtract Mechanism with joint spatialtemporal analysis captures complementary dynamic incoherence, which aids detection. However, the approach remains sensitive to compression and demands substantial computing resources, impacting its scalability. Collectively, these methods present valuable insights and advancements in deepfake detection. Yet, recurring challenges—including computational costs, sensitivity to video compression, and overfitting in constrained datasets—highlight areas for future research. Enhancements in efficiency, generalization, and scalability remain critical to achieving robust, deployable deepfake detection systems.

3. Proposed Method

EfficientNet for Image Feature Extraction EfficientNet, a CNN architecture optimized for performance and computational efficiency, is employed for extracting geometric features from face landmarks. This model improves upon traditional CNNs by providing better accuracy and requiring fewer resources. Transformer for Temporal Dependency Modeling Transformers, widely used in natural language processing, have shown superiority in capturing complex temporal relationships. We replace GRU with a Transformer model to improve the temporal feature analysis of sequential face landmarks. This allows the model to recognize patterns over long sequences, enhancing its ability to detect fake content effectively. Ensemble learning: Ensemble learning improves deepfake detection by combining spatial (frame-level) and temporal (motion-based) features for robust classification. We use EfficientNetB0 for spatial feature extraction and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

TimeSformer for temporal patterns, fusing their outputs for final prediction. A Multi-Layer Perceptron (MLP) classifier processes the combined features, while a soft voting ensemble enhances accuracy. This approach ensures better detection of both visual and motion inconsistencies, reducing false positives and negatives. [1]

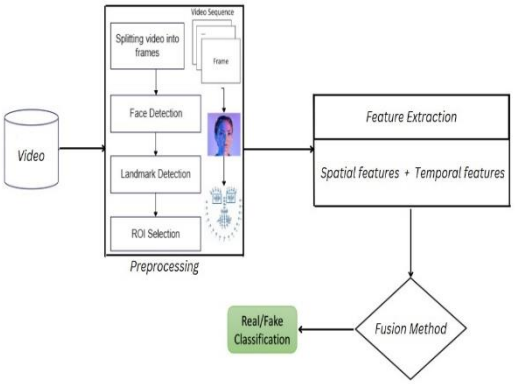

Figure 2 Model Architecture

4. Mathematical Formulation

4.1.Input Data

The input video is represented as a sequence of frames:

$V = \{F_1, F_2, ..., F_N\}$ (1)

where F_i denotes the i^{th} frame of the video, and N is the total number of frames. [3]

4.2.Facial Landmark Extraction

From each frame, key facial landmarks are extracted:

$L_i = \{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$ (2) where m is the number of landmark points per frame.

Spatial Feature Extraction Using EfficientNet

EfficientNet is used to extract deep visual features from each frame.

$Ei = f_{EfficientNetB3}(Fi)$ (3)

4.3.Temporal Feature Analysis Using Transformer

A Transformer model analyzes the temporal evolution of facial landmarks:

$T = fTransformer(\{L1, L2, ..., LN\})$ (4)

4.4.Feature Fusion

The extracted spatial and temporal features are fused using a weighted sum:

$Fcombined = wE \cdot E + wT \cdot T$ (5)

where w_E and w_T are the respective weights for spatial and temporal features, and E is the aggregated spatial feature vector. (Figure 3) [2]

4.5.Classification

The combined feature vector is passed through a sigmoid activation to obtain a classification probability:

$D = \sigma(fclassify(Fcombined)) = \frac{1}{1+e^{-f_{classify}(F_{combined})}}$ (6)

Final Decision

The final decision is based on a threshold applied to the output probability D:

$$Decision = \begin{cases} 1, & \text{if } D \geq 0.5 \text{ (Real)} \\ 0, & \text{if } D < 0.5 \text{ (Fake)} \end{cases}$$ (7)

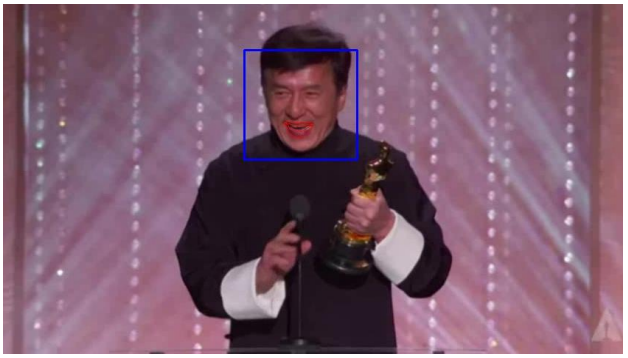

Figure 3 ROI Detection

5. Image-Based Deepfake Detection

Advancements in convolutional neural network (CNN)based architectures, such as VGG, ResNet, and EfficientNet, have significantly improved image-based deepfake detection. VGG and ResNet are widely used due to their strong feature extraction capabilities, but they often suffer from high computational costs. EfficientNet addresses these limitations by employing compound scaling, which optimally balances depth, width, and resolution to achieve better accuracy with fewer parameters. This scaling method ensures improved performance while reducing the computational burden, making EfficientNet an ideal choice for resource-efficient applications. [4]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

6. Sequential Modeling in Deepfake Detection

Traditional methods for modeling temporal dependencies in deepfake detection include recurrent neural networks (RNNs) and gated recurrent units (GRUs). These models excel at capturing short-term dependencies but struggle with longterm patterns due to issues like vanishing gradients. Transformers, on the other hand, use self-attention mechanisms to model dependencies across the entire sequence, eliminating the reliance on sequential processing. This makes Transformers more effective for handling complex temporal relationships in video-based deepfake detection. [5]

7. Multimodal Deepfake Detection Frameworks

Multimodal frameworks, such as the Face and Audio-based Multimodal (FAMM) model, combine visual and audio cues to improve deepfake detection accuracy. These frameworks leverage complementary information from different modalities, enhancing robustness against adversarial attacks and reducing errors associated with unimodal approaches. By integrating audio and video features, multimodal systems can better capture inconsistencies in synthetic media, leading to improved performance in real-world scenarios.

8. Enhancements in Model Design

8.1.Advantages of EfficientNet

EfficientNet offers several advantages over traditional CNN architectures:

- **Efficient Scaling**: Compound scaling balances depth, width, and resolution, enabling better trade-offs between computational cost and accuracy.
- **Transfer Learning:** Pre-trained weights on large datasets like ImageNet allow for faster training and improved generalization.
- **Geometric Feature Handling:** EfficientNet can effectively process facial landmarks using geometric transformations for enhanced feature extraction.

8.2.Advantages of Transformers

Transformers bring significant benefits for temporal modeling:

- **Self-Attention Mechanisms:** These mechanisms enable the model to focus on relevant temporal relationships across sequences.
- **Scalability:** Transformers can handle large datasets efficiently due to their parallel processing capabilities.
- **Non-Sequential Processing:** Unlike GRUs or RNNs, Transformers do not rely on sequential processing, avoiding issues like vanishing gradients and enabling better long-range dependency modeling. [6]

8.3.Fusion Techniques

Fusion techniques combine the outputs of multiple models to improve overall prediction accuracy:

- **Comparison of Methods:** Weighted fusion, stacking, concatenation, and soft voting are common methods for integrating model predictions.
- **Advantages of Weighted Fusion:** By dynamically assigning importance to models based on their performance, weighted fusion ensures a balanced and accurate final prediction. [7]

9. Dataset

We used the FaceForensics++ dataset available on Kaggle.https://www.kaggle.com/datasets/hungle3401/faceforensicsDataset Link In this context, the FaceForensics++ (FF+) dataset is particularly valuable, as it contains a diverse collection of deepfake videos generated using multiple methods, such as the popular DeepFake, FaceSwap, and AutoEncoder techniques. By training deepfake detection models on datasets like FF+, the model is exposed to a wide spectrum of manipulation techniques, enhancing its ability to detect deepfakes in real-world applications, where new generation techniques may emerge. [8]

10. Results and Evaluation

We evaluated our approach using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

We implemented our model using PyTorch. For the Efficient- Net component, we utilized EfficientNet-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

B0 pretrained on ImageNet and fine-tuned it on our facial landmark data. The Transformer architecture consisted of 4 encoder layers with 8 attention heads and a hidden dimension of 512. We trained the model using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. (Table 1) [9]

Table 1 Classification Report for Real and Fake Classes

Class	Precision	Recall	F1-Score	Support
Real	0.85	0.94	0.89	35
Fake	0.91	0.77	0.83	26
Accuracy			0.87	61
Macro Avg	0.88	0.86	0.86	61
Weighted Avg	0.87	0.87	0.87	61

Despite advancements in deepfake detection, several challenges continue to hinder real-world deployment. One significant limitation is the scarcity of diverse, large-scale datasets, which often leads to class imbalance and biased model predictions. While advanced models such as EfficientNet and Transformers demonstrate high accuracy, they require substantial computational resources, which constrains their applicability in real-time or resource-limited environments. These deep learning models are also vulnerable to adversarial attacks—minor, imperceptible alterations to inputs that can deceive the system—raising concerns about robustness and reliability. Scalability remains a challenge, especially when processing high-resolution videos or long temporal sequences, both of which increase computational overhead. Furthermore, the opaque nature of these AI models (often described as "black boxes") hinders interpretability, reducing trust in critical or sensitive applications. The fast-paced evolution of deepfake generation techniques also presents an ongoing challenge, as detection systems struggle to adapt. Finally, deploying these systems raises ethical concerns regarding privacy, consent, and potential misuse, particularly when applied in surveillance or legal contexts.

Conclusion

Deepfake detection has emerged as a critical area of research due to the rapid proliferation of synthetic media in various domains. This paper highlighted the integration of advanced architectures like EfficientNet for robust feature extraction and Transformers for superior temporal modeling, as well as the advantages of multimodal frameworks and weighted fusion techniques for improving detection accuracy. Despite these advancements, challenges such as dataset limitations, model complexity, and adversarial vulnerabilities remain. Future research should focus on enhancing multimodal frameworks by integrating additional modalities like audio or physiological signals, developing lightweight architectures for real-time applications, and implementing adaptive learning systems to address emerging deepfake techniques. Furthermore, increasing the interpretability of detection models through explainable AI, improving adversarial robustness, and expanding scalable datasets with diverse samples are vital steps toward building robust, transparent, and effective deepfake detection systems that can meet the demands of realworld scenarios. [10]

Acknowledgment

We sincerely thank our guide, Ms. Hema Jadhav, for her invaluable guidance and continuous support throughout this project. Her insights and expertise were instrumental in the successful completion of our work. We are also grateful to our project coordinator, Ms. Rohini Naik, for her constant motivation, encouragement, and assistance during the project. We extend our heartfelt gratitude to Dr. C. S. Kulkarni, Head of the Department of AIDS, and Dr.Sudhir.B.Lande, Principal of VPKBIET, Baramati, for providing us with the necessary resources and a conducive environment for research and development. Finally, we express our appreciation to the teaching staff of the Department of AIDS for their encouragement, constructive feedback, and unwavering support, which significantly contributed to the success of this work.

References

[1]. Yumei Wang," FAMM: facial muscle motions for detecting compressed deepfake videos over social networks," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, Dec. 2023.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2268 - 2273
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0334

[2]. Juan Hu,"Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, Mar. 2022.Guilin Pang, "MRE-Net: multi-rate excitation network for deepfake video detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, Aug. 2023.

[3]. Akash Chintha, "Recurrent convolutional structures for audio spoof and video deepfake detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, Mar. 2022.

[4]. Lulu Tian, "FakePoI: a large-scale fake person of interest video detection benchmark and a strong baseline," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, Mar. 2023.

[5]. Yang Yu, "ISTVT: interpretable spatial-temporal video transformer for deepfake detection," IEEE Transactions on Information Forensics and Security, vol. 18, 2023.

[6]. Cairong Zhao, "MSVT: multiple spatiotemporal views transformer for deepfake video detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, Sep. 2023.

[7]. Li Zhang, "Unsupervised learning-based framework for deepfake video detection," IEEE Transactions on Multimedia, vol. 25, 2023.

[8]. Wang, H., "Exploiting Complementary Dynamic Incoherence for DeepFake Video Detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, 2023.

[9]. Ramadhani, K. N., "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution, and Self Attention," IEEE Access, vol. 12, 2024.

[10]. Mingxing Tan and Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the International Conference on Machine Learning (ICML), 2019.