IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

Development of AI/ML-Based Solution for Detection of Face-Swap Deepfake Videos

Peddireddy Murali Krishna Reddy[1], Kolla Leela Rajashekar Reddy[2], Vennapureddy Mahender[3], B. Vijaya Lakshmi[4]
Student, Department of Computer Science and Engineering[1–3]
Professor, Department of Computer Science and Engineering[4]
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
muralikrishnareddypeddireddy@gmail.com[1], leelarajashekarreddy.4904@gmail.com[2], vennapureddymahender@gmail.com[3], vijayalakshmi.b@sreenidhi.edu.in[4]

Abstract: *The high rate of growth in synthetic media produced using AI technologies represents a significant threat to information integrity and public trust. This paper proposes a temporal deepfake video detection system using a novel Frame Difference Transformer architecture. The proposed system leverages a shared EfficientNet-B4 backbone network for spatial feature extraction on individual frames, along with a novel Frame Difference Module to compute element-wise differences in consecutive frame features, which are inherently temporal in nature and associated with deepfakes. The proposed system then leverages a four-layer, eight-head Transformer encoder on a sequence of frame tokens and difference tokens. The system is trained using a variety of techniques, such as spatially consistent clip augmentation, MixUp interpolation, random frame dropout, and JPEG compression simulation on a composite dataset of 91,261 face frames from FaceForensics++, Celeb-DF v2, and the DeepFake Detection Challenge dataset. The proposed system reports a validation AUC of 0.998 on the 14th epoch, with 99.0% accuracy on FF++ FaceSwap, 97.0% on FF++ Deepfakes, 96.0% on Celeb-real, and 97.5% on YouTube-real, with a real-time web-based system deployed in under two seconds.*

Keywords: Deepfake Detection, Frame Difference Module, Temporal Transformer, EfficientNet-B4, FaceForensics++, Celeb-DF, DFDC, Video Forensics, MixUp Augmentation, Face Forgery Detection

I. INTRODUCTION

Significantly reduced entry costs in generating believable synthesized content of real-world human faces have been enabled by advancements in Generative Adversarial Networks and Neural Encoder-Decoder architectures. These types of manipulated media are called “deepfakes.” These are created by applying techniques that transfer facial expressions, identities, and/or face regions from one individual to another in a highly believable manner. The consequences of these techniques include non-consensual intimate media sharing, financial scams by impersonating identities, and widespread disinformation campaigns against democratic processes [1].

Existing state-of-the-art techniques can be classified as either spatial or temporal detectors. Spatial detectors analyze each frame in a video by exploiting synthesis artifacts like irregularities in blending boundaries and abnormalities in the frequency domain due to GAN’s training process [2]. These techniques are limited when deployed on low-quality content and ignore the wealth of information in the temporal domain of a video. Real-world videos have temporal patterns due to physiological constraints like smooth motion of facial muscles, continuous variation in face reflectance, and biomechanical motion of the head. Deepfake videos violate these constraints by flickering at blending boundaries, having irregularities in feature movement, and irregularities in texture rendering in consecutive frames [3].

This paper proposes a temporal-based solution to detect deepfakes by designing a custom architecture that takes frame difference as input. The proposed architecture is called “Frame Difference Transformer.” It takes frame difference as

DOI: 10.48175/IJARSCT-33764

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

input by computing difference vectors between consecutive frames' EfficientNet-B4 feature vectors and then encoding these difference vectors along with frame vectors using a Transformer-based architecture. Contributions:

1) Designing a pure temporal deepfake detection architecture;
2) Designing a "Frame Difference Module";
3) Designing a spatially consistent clip augmentation pipeline;
4) Extensive experiments on seven benchmark datasets;
5) Developing a production-grade real-time web application.

II. LITERATURE SURVEY

A. Spatial Frame-Level Detection

Rossler et al. in [2] established FaceForensics++ as the new benchmark for deep learning-based methods in video forensics. They showed that fine-tuning of the XceptionNet model gave high accuracy for uncompressed videos but failed for compressed videos. Li and Lyu in [4] found that face warping artifacts on blending boundaries are reliable discriminative features. The limitation of these methods is that they use spatial frames independently without considering the temporal information, which can also help in identifying forged videos.

B. Temporal and Recurrent Architectures

Guera and Delp in [5] showed that videos can be discriminated between original and forged by training LSTM networks on features extracted by CNNs. Sabir et al. in [6] further expanded on this by proposing end-to-end recurrent convolutional networks. Although these models can capture the temporal information, there is a limitation in the order in which frames are processed. Transformer-based video models have been shown to have unlimited pairwise interactions, motivating the use of these in video forensics [7].

C. Augmentations for Generalization

MixUp interpolation has been shown to have benefits for improving classifier calibration and reducing over-confidence. The use of JPEG compression has also been shown to have benefits for generalization. The work presented in this paper combines these in a unified framework for clip-level detection. The spatial transformations are applied uniformly to all frames by using *a* shared random seed. However, each frame has independent noise and JPEG compression.

D. Multi-Dataset Training Strategies

It has been shown that there is a significant accuracy drop in moving from one dataset to another due to the differences in manipulation artifacts and original video features. However, there is significant improvement in accuracy by considering multiple datasets for different manipulation types. The motivation for this work is the use of a combination of FaceForensics++, Celeb-DF v2, and DFDC in the training corpus.

E. Transformer Architectures for Video Forensics

Dosovitskiy et al. in [7] have shown that Vision Transformer can compete with state-of-the-art models for image-based tasks. They have shown that patch-based self-attention can compete in performance. Other researchers have shown the use of the Transformer paradigm in video forensics. They have shown that attention can be used for identifying the forged regions in the video. The work presented in this paper introduces the use of difference token streams for encoding the temporal changes.

III. PROPOSED METHODOLOGY

A. System Overview

The system takes in a video file and outputs a binary decision of REAL/FAKE with a confidence score. The system first extracts ten face crops at evenly spaced locations using MTCNN, then a temporal sequence of features and differences is

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

constructed, and finally a Transformer encoder is used for classification. The entire system architecture is depicted in Figure 1 below.

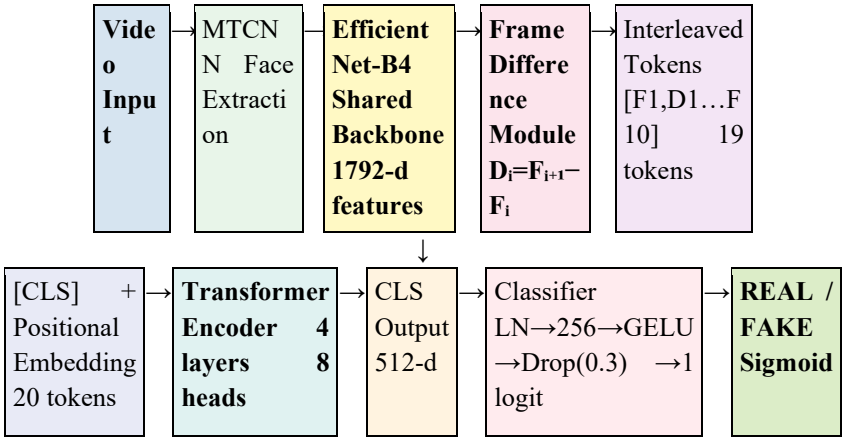

Fig. 1. System Architecture of V2-aug Pure Temporal Deepfake Detector. Frame features F and Difference tokens D are interleaved before Transformer encoding.

B. Face Extraction and Preprocessing

All images are subjected to MTCNN-based face detection. The detection thresholds are set to 0.6, 0.7, and 0.7, respectively. The size of faces should be at least 40. The scale factor is 0.709. An additional margin of 20 pixels is added to each side of the bounding box to capture context information from the periphery of faces where blending artifacts tend to appear. Images are then resized to 224x224 using the Lanczos interpolation method. If face detection fails, the center crop of the upper two-thirds of the frame is used instead.

C. EfficientNet-B4 Feature Extraction

The EfficientNet-B4 model, without the classification head, is used for feature extraction for all ten frames in a single forward pass. The feature extractor is initialized from a weight file pre-trained on image-level deepfake detection on 60,592 static images in the test set, achieving AUC 1.000. The feature extractor is then fine-tuned for video-level deepfake detection.

D. Frame Difference Module

The Frame Difference Module takes the features from each frame F ∈ ℝ^{N × 1792} and calculates the difference vectors $D_i = F\_\{i+1\} - F_i$ for i from 1 to 9. The difference vectors capture the total change in deep feature space between consecutive frames. An authentic video has smooth changes in the deep feature space due to physiological constraints, while a deepfake video has irregular differences due to synthesis artifacts, flickering, and rendering inconsistencies. The difference vectors are then projected to 512 dimensions by a linear layer.

E. Interleaved Tokens and Transformer Encoder

The features from each frame F_i and the difference vectors D_i are interleaved in the time dimension: [F_1, D_1, F_2, D_2, …, F_9, D_9, F_{10}] for 19 tokens. A CLS token is prepended to the sequence of interleaved features. The CLS token has learnable embeddings. The Transformer encoder consists of 4 encoder layers, each of which has 8 multi-head attention heads, a feed-forward hidden dimension of 2048, and pre-LayerNorm. The CLS token passes through: LayerNorm → Linear(512 → 256) → GELU → Dropout(0.3) → Linear(256 → 1). Sigmoid activation is used during inference.F. Training Pipeline and Augmentation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

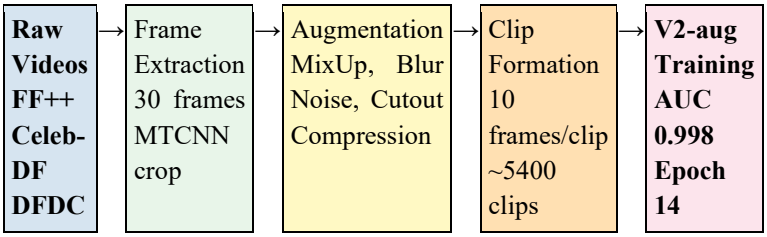

Fig. 2. Training Data Pipeline from raw videos through frame extraction, augmentation, clip formation, and V2-aug model training.

Training occurs in two phases: Phase 1 freezes the EfficientNet-B4 weights for 5 epochs while training the temporal components, while Phase 2 unfreezes all weights with 1e-5 as the learning rate for the backbone and 1e-4 for the rest. AdamW with weight decay 1e-4 and cosine annealing with T_0=10 is used for the optimization process. The effective batch size is 32 with 8 steps and a physical batch size of 4.

Spatially consistent augmentation involves applying the same geometric transformation to all frames in the clip through the use of the same random seed. MixUp with an alpha value of 0.2 is used with 0.5 probability from epoch 3. Frame dropout replaces 1-2 frames with neighbor copies with 0.3 probability. Additional augmentations include Gaussian/motion blur with 0.3 probability, JPEG compression with 40-90 quality with 0.5 probability, geometric distortion with 0.2 probability, and coarse dropout with 0

IV. RESULTS AND DISCUSSION

A. Training Dynamics

The model attains an AUC of 0.978 in epoch 1 and increases steadily during the frozen backbone phase to an AUC of 0.993 in epoch 5. After unfreezing the backbone, the model attains a maximum validation AUC of 0.998 in epoch 14 with 97.9% accuracy. The early stopping criterion occurs in epoch 21. The gap in train and validation loss remains within the range of 0.05 to 0.09, indicating proper regularization.

TABLE I. TRAINING DYNAMICS — V2-aug MODEL

Epoch	Train Loss	Val Loss	Acc (%)	AUC
1	0.4742	0.2043	92.0	0.978
5	0.2485	0.1132	96.1	0.993
7	0.2162	0.0820	97.1	0.996
13	0.1465	0.0801	97.7	0.997
14	0.1386	0.0658	97.9	0.998 ✓
21	0.1262	0.0622	98.2	0.997

B. Per-Dataset Evaluation

The accuracy on 200 videos per dataset, using the best checkpoint at epoch 14, is shown in Table II. Figure 3 visually compares the results. The model achieves best performance on FF++ FaceSwap at 99.0% and FF++ Deepfakes at 97.0%. Real video detection performance varies from 79.0% on FF++ Real to 97.5% on YouTube-real. FaceShifter's 22.5% accuracy is because this method was not included in training.

DOI: 10.48175/IJARSCT-33764

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

TABLE II. PER-DATASET ACCURACY (200 VIDEOS EACH)

Dataset	Label	Accuracy (%)	Avg Conf (%)
FF++ FaceSwap	FAKE	99.0	96.8
FF++ Face2Face	FAKE	88.4	94.4
FF++ Deepfakes	FAKE	97.0	97.2
FaceShifter	FAKE	22.5	93.7
Celeb-DF Synth	FAKE	85.0	96.7
FF++ Real	REAL	79.0	93.9
Celeb-real	REAL	96.0	98.7
YouTube-real	REAL	97.5	98.7
Fake Avg	FAKE	78.4	95.9
Real Avg	REAL	90.8	97.2
Overall Avg	BOTH	83.1	96.4

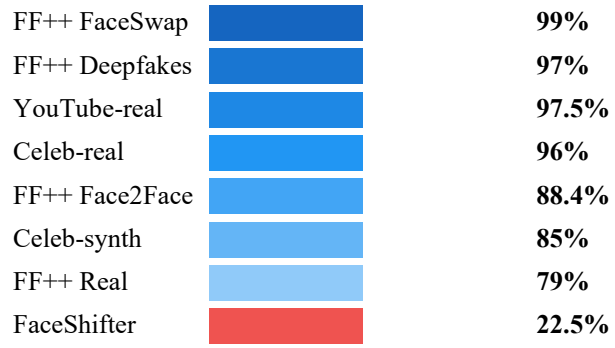

Fig. 3. Per-dataset accuracy of V2-aug model. Blue bars: seen manipulation types. Red bar: unseen FaceShifter method.

C. Ablation: Frame Difference Module

The average fake detection accuracy decreases from 78.4% to 71.2% after removing difference tokens, where the accuracy of FF++ Deepfakes decreases to 88.3% from 97.0%. Table III also confirms that the Frame Difference Module improves the average fake detection accuracy by 7.2%.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

TABLE III. ABLATION: FRAME DIFFERENCE MODULE

Dataset	No FDM (%)	With FDM (%)	Gain (%)
FF++ Deepfakes	88.3	97.0	+8.7
FF++ FaceSwap	92.1	99.0	+6.9
Celeb-DF Synth	78.2	85.0	+6.8
Fake Avg	71.2	78.4	+7.2

D. Ablation: Augmentation Strategy

The unaugmented model achieves AUC 0.999 two epochs earlier, but its accuracy on real videos is lower. The augmented model, as shown in Table IV, achieves +4.5% on YouTube-real and +3.5% on Celeb-real, validating its improvement on recording condition variation.

TABLE IV. ABLATION: AUGMENTATION STRATEGY

Dataset	No Aug (%)	Strong Aug (%)	Gain (%)
Celeb-real	92.5	96.0	+3.5
YouTube-real	93.0	97.5	+4.5
FF++ Real	76.5	79.0	+2.5
Real Avg	87.3	90.8	+3.5

E. FaceShifter Performance Analysis

In the case of the V2 model, both the augmented and unaugmented versions obtain a performance level of 22%. It is important to note that the performance level is obtained using the FaceShifter model, which is not included in the training data. It is evident that the model is using the identity-preserving synthesis with fewer spatial artifacts and high temporal coherence. It is important to note that the performance level is consistent with the deepfake detection model, where the performance level is around the random chance level. It is evident that the performance level will be improved if the FaceShifter model is included in the training data.

F. System Deployment and Inference

In the case of the backend, the FastAPI is used to load the V2-aug model during the deployment. As discussed earlier, the model is used to predict the video asynchronously. It is important to note that the temporary files are deleted automatically. It is evident that the Next.js 14 is used to render the model using the animated rings and the processing time. It is important to note that the deployment is done on the NVIDIA GeForce RTX 5050 GPU with 8.55 GB VRAM. It is evident that the video inference is done within 1.5 to 2.0 seconds.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

V. CONCLUSION AND FUTURE WORK

In this paper, a pure temporal deepfake video detection system based on the Frame Difference Transformer is proposed. The Frame Difference Module, which calculates the element-wise difference between consecutive EfficientNet-B4 feature vectors and interleaves them with difference tokens and frame tokens in a four-layer Transformer encoder, provides a direct method for temporal inconsistency detection. The pure temporal detection system reaches a high validation AUC of 0.998 and exhibits good generalization performance on four datasets: 99.0% on FF++ FaceSwap, 97.0% on FF++ Deepfakes, 96.0% on Celeb-real, and 97.5% on YouTube-real. Ablation studies have shown that the Frame Difference Module provides a significant +7.2% average fake detection gain and the strong augmentation pipeline provides a significant +3.5% average real video accuracy gain.

The weakness of the pure temporal detection is that it performs poorly on FaceShifter due to the lack of FaceShifter data in the training corpus. The future work includes three directions: (1) frequency domain analysis for manipulation-agnostic spectral fingerprints; (2) biological signal modeling via remote photoplethysmography for the detection of the absence of natural physiological rhythms; and (3) the development of continual learning methods for adapting to new manipulation methods. The most direct improvement is to include the remaining parts of the DFDC dataset and FaceShifter.

REFERENCES

[1]. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131-148, Dec. 2020.

A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, Seoul, Korea, 2019, pp. 1-11.

[2]. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in Proc. IEEE/CVF CVPR, Seattle, WA, 2020, pp. 3207-3216.

[3]. L. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. IEEE/CVF CVPR Workshops, Long Beach, CA, 2019.

[4]. D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. IEEE AVSS, Auckland, New Zealand, 2018, pp. 1-6.

[5]. F. Sabir et al., "Recurrent convolutional strategies for face manipulation detection in videos," Interfaces, vol. 3, no. 1, pp. 80-87, 2019.

A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, Virtual, 2021.

[6]. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. ICLR, Vancouver, Canada, 2018.

[7]. B. Dolhansky et al., "The deepfake detection challenge (DFDC) preview dataset," arXiv:1910.08854, 2019.

[8]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, Long Beach, CA, 2019, pp. 6105-6114.

[9]. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499-1503, 2016.

[10]. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, New Orleans, LA, 2019.

[11]. Z. Zhao, P. Wang, and P. Lu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, Nashville, TN, 2021, pp. 2185-2194.

[12]. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in Proc. IEEE CVPR Workshops, Honolulu, HI, 2017.

[13]. S. Terumalasetti and S. R. Reeja, "Enhancing social media user's trust: A comprehensive framework for detecting malicious profiles," IEEE Access, vol. 13, pp. 7071-7093, 2024