Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062954

ARTICLE

Real-Time Deepfake Detection via Gaze and Blink Patterns: A Transformer Framework

Muhammad Javed[1], Zhaohui Zhang[1,*], Fida Hussain Dahri[2], Asif Ali Laghari[3,*], Martin Krajčík[4] and Ahmad Almadhor[5]

[1]Department of Computer Science and Technology, College of Computer Science, Donghua University, Shanghai, 200022, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China
[3]Software College, Shenyang Normal University, Shenyang, 110136, China
[4]Department of Information Management and Business Systems, Faculty of Management, Comenius University, Bratislava Odbojárov 10, Bratislava, 82005, Slovakia
[5]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, 72388, Saudi Arabia

*Corresponding Authors: Zhaohui Zhang. Email: zhzhang@dhu.edu.cn; Asif Ali Laghari. Email: asiflaghari@synu.edu.cn

Received: 31 December 2024; Accepted: 06 June 2025; Published: 29 August 2025

ABSTRACT: Recent advances in artificial intelligence and the availability of large-scale benchmarks have made deepfake video generation and manipulation easier. Therefore, developing reliable and robust deepfake video detection mechanisms is paramount. This research introduces a novel real-time deepfake video detection framework by analyzing gaze and blink patterns, addressing the spatial-temporal challenges unique to gaze and blink anomalies using the TimeSformer and hybrid Transformer-CNN models. The TimeSformer architecture leverages spatial-temporal attention mechanisms to capture fine-grained blinking intervals and gaze direction anomalies. Compared to state-of-the-art traditional convolutional models like MesoNet and EfficientNet, which primarily focus on global facial features, our approach emphasizes localized eye-region analysis, significantly enhancing detection accuracy. We evaluate our framework on four standard datasets: FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb. The proposed framework results reveal higher accuracy, with the TimeSformer model achieving accuracies of 97.5%, 96.3%, 95.8%, and 97.1%, and with the hybrid Transformer-CNN model demonstrating accuracies of 92.8%, 91.5%, 90.9%, and 93.2%, on FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb datasets, respectively, showing robustness in distinguishing manipulated from authentic videos. Our research provides a robust state-of-the-art framework for real-time deepfake video detection. This novel study significantly contributes to video forensics, presenting scalable and accurate real-world application solutions.

KEYWORDS: Deepfake detection; deep learning; video forensics; gaze and blink patterns; transformers; TimeSformer; MesoNet4

1 Introduction

Deepfake media increased rapidly due to the momentum gained from recent AI, deep learning, and generative model achievements, which seriously affected the authenticity and security of media [1,2]. In the security sector, there is a need for effective deepfake detection mechanisms that could reduce risks of identity theft, fraud, and misinformation that could deplete public confidence [1–3].

Deepfakes use generative models, such as Autoencoders (AEs) [3], Generative Adversarial Networks (GANs) [4], Neural Radiance Fields (NeRFs) [5], and Diffusion Models [6], to produce realistic manipulated content like videos [7], images [8], and audio that convincingly mislead viewers with great accuracy [7,9]. Free access to advanced generative algorithms such as AEs [3], GANs [4], and large public datasets has accelerated the accessibility of deepfake creation tools and datasets. It has significantly amplified the challenges surrounding their detection. Sophisticated algorithms, coupled with widely available computational resources, have drastically reduced the technical barriers to producing compelling content [10–12], and compelling fake videos, making the process faster and easier [13–15]. Consequently, deepfake technology has improved dramatically, resulting in a substantial increase in the volume of spoofed content, which often cannot be detected even by human observers. While these technologies have potential applications in entertainment and education, for example, their misuse has substantial concerns: the spread of misinformation, damage to reputations, and threats to societal trust and digital integrity [7,10,12,16].

Usually, there are four major classes of deepfake video generation techniques [17]: reenactment (1), swapping (2), editing (3), and synthesis (4). Face reenactment [18] involves the synthesis of an individual's facial expressions and movements in a target video to mimic those of a person from a source video [19]. On the other hand, face swapping [20] or replacing refers to the complete replacement of the face of the individual in the target video with another person's face from the source video. Deepfakes of the editing type include modifications of particular features in a target video, such as face, age, weight, ethnicity, or any other characteristics, through addition, modification, or removal. Synthesis [21,22]: Deepfakes can be created without having any specific target for reference, enabling one to build completely artificial persons or characters, as used in online media or for creative applications in movies and games. So, whereas the former two methods mainly concern the manipulation of facial expressions, the latter involves complete identity change.

Recent detection methods and approaches primarily targeted visual artefacts, such as irregular eye colour, absent reflections, or inconsistencies in facial details, which were prevalent in the initial generations of deepfake videos [23–27]. Furthermore, the fast progress in AEs and GANs has raised the realism and quality of fake content, rendering these artefact-based approaches largely ineffective [28,29]. These studies have indeed tended towards deep learning-based methods using large-scale public datasets, such as FaceForensics++ [11], Celeb-DF [15], and DFDC-preview [30], to overcome these limitations. Most of these methods use convolutional neural networks (CNNs) to extract spatiotemporal features for forged content detection [25,31,32]. Temporal methods often use RNNs to check for temporal sequences showing inconsistencies and disruptions between frames [31]. Spatial artefact detection, in turn, concentrates on frame-by-frame analysis using Convolutional Neural Network (CNN) architectures such as VGG16 [31], ResNet50 [33], and Xception [33]. Although these approaches have enjoyed worthy results, this reliance on specific data set patterns often bottlenecks generalization to unseen data or novel manipulation techniques.

Most existing approaches focus on global facial features [34–36]. They thus may fail to take full advantage of the localized, dynamic cues, such as gaze direction and blink patterns, which have been essential in finding minor inconsistencies in synthetic media. Current studies and state-of-the-art work in deepfake detection and face forgery, such as CNN and hybrid deep learning models [37], underscore the advantage of analyzing localized facial features. These often lack the time sensitivity to easily pick out minor variations in gaze trajectories or blink intervals of interest to incorporate these limitations, developing a novel framework incorporating spatial and temporal analysis dimensions, focusing on finer-level features for improved robustness and generalization. Therefore, gaze and blink patterns provide an encouraging way to detect deepfakes robustly [38]. Natural eye movement-characterized by spontaneous blinks and shifts in gaze direction is hard to imitate in believable manipulated videos owing to the natural intricacies in human

physiology and behaviour [2]. Therefore, our proposed framework overcomes these limitations of recent studies and finds an essential place in journalism and content creation for verifying video content to ensure the audience can rely on the media they consume. Social networking sites will also be able to flag and manage manipulated content in real-time by using deepfake detection technologies, making them safer. Because of this, the research will develop deepfake detection from a technical perspective and indicate the vibrant need for such inventions in modern society.

To overcome these issues and limitations, we proposed a novel real-time deepfake video detection framework focusing on fine-grained features, such as gaze and blink pattern analysis [38]. This novel proposed framework leverages the strength of the transformer-based model [39,40], the TimeSformer model, which surpasses in modelling spatio-temporal patterns in video content [40,41]. The TimeSformer provides robust spatiotemporal attention mechanisms. We extend its applicability by incorporating fine-grained gaze and blink pattern analysis, which has not been previously explored in the context of deepfake detection. Our framework tailors TimeSformer to focus on localized eye-region features, addressing specific challenges in real-time manipulation detection. The TimeSformer model, initially proposed by Facebook in 2021, has demonstrated exceptional performance in general video classification tasks [42,43]. However, its potential for addressing fine-grained manipulations in facial regions, such as gaze shifts and blink anomalies, remains undiscovered. This study builds upon the TimeSformer framework by introducing a novel application tailored for real-time deepfake detection, leveraging eye-region-focused spatiotemporal attention. To achieve this, we incorporated precise Region of Interest (ROI) extraction, temporal smoothing of eye dynamics, and custom preprocessing, enabling TimeSformer to capture subtle forgery indicators unique to manipulated videos, which represents a significant advancement in applying transformer-based models for deepfake detection. We include another hybrid Transformer-CNN model as a baseline for comparative analysis, enabling a thorough evaluation of the trade-off detection accuracy. The hybrid Transformer-CNN architecture leverages the lightweight spatial feature extraction capabilities of MesoNet4 alongside the robust temporal modelling of TimeSformer. While existing studies often utilize CNNs or Transformers in isolation, our framework integrates the two to improve computational efficiency while maintaining high detection accuracy for fine-grained eye-region manipulations. Our novel framework is evaluated on four standard datasets: FaceForensics [11], CelebDF-V2 [15], DFDC [30], and FakeAVCeleb [44]. This novel study presents a transformative framework for real-time deepfake video detection through the following contributions:

- A novel adaptation of the TimeSformer model for deepfake detection, focusing on gaze and blink patterns through customized preprocessing, including ROI extraction and temporal smoothing, to address the unique challenges of video spatio-temporal manipulation.
- Developing a Transformer-CNN hybrid framework to analyze the trade-offs between lightweight CNN architectures and transformer-based models in deepfake detection.
- Rigorous testing on benchmark datasets (FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb) focuses on Intra and Cross-Dataset Evaluation, ensuring the framework's robustness.
- To optimize the framework for real-time deployment, achieving high accuracy and computational efficiency suitable for practical applications.

The paper is organized as follows: Section 2 reviews the literature on deepfake detection techniques, while Section 3 presents a complete methodology of the proposed framework, including preprocessing, model architectures, training, and evaluation metrics. Section 4 discusses the analysis of the experimental results and discussion, including comparative evaluations. Finally, Section 5 concludes the paper with key findings and possible directions for future research.

2 Literature Review

Recent studies have shown a significant advancement in real-time deepfake video detection and face forgery detection [1,2,36,45,46]. The misuse of deepfake technology has regularly attracted extensive attention, and various approaches have been proposed in-depth to detect deepfake videos, synthetic media, and face forgery to overcome the deepfake challenges. This literature review briefly summarises the familiar and recent state-of-the-art work of deepfake approaches and strategies for face forgery and deepfake video detection. In a study [47], researchers introduced a system for detecting deepfakes by leveraging the temporal information within video streams to identify inconsistencies across multiple frames in manipulated videos. To analyze temporal data, the researchers utilized a recurrent convolutional framework [47], which combines a CNN for feature extraction with a Bidirectional Recurrent Neural Network (BiDir RNN) to process temporal patterns in video sequences. They specifically explored two CNN architectures, ResNet [48] and DenseNet [49], for extracting features and included a detailed preprocessing strategy for preparing facial frames before feeding them into the models. The performance of the models was assessed using the widely recognized FaceForensics++ deepfake detection benchmark [11], demonstrating solid results in an intra-dataset evaluation. However, the study did not include a cross-dataset analysis.

The study [50] deviated from conventional image features by utilizing biological signals, such as photoplethysmography (PPG), which captures subtle changes in colour and motion in RGB videos, to develop their model. The approach integrated a Convolutional Neural Network (CNN) with a Support Vector Machine (SVM). Each model independently classified features, and their outputs were combined to generate a final classification score. This deepfake detection method demonstrated strong performance across intra-dataset and inter-dataset testing scenarios, evaluated on various deepfake detection benchmarks, including Celeb-DF [15], FaceForensics [11], and FakeAVCeleb [44]. The study by [51] introduced a compact deep architecture named "Audio-Visual Person-of-Interest DeepFake Detection", which enhances entirely generated facial image detection by sharing localized features across altered regions. In 2022, a study by [52] proposed combining EfficientNetB0 convolution and various Vision Transformers (ViTs), achieving effective video deepfake detection through a straightforward voting mechanism. Another study [53], in which researchers investigated a hybrid model integrating convolutional neural networks (CNNs) with visual Transformers (ViTs), produced competitive results on the DFDC dataset. Meanwhile, another study [54] introduced an unsupervised CNN architecture that integrated 3D-CAE and 3D-CNN, creating a deep representation framework (Xi) to leverage unsupervised and supervised 3D image input analysis techniques. A study [55] in which they developed an interpretable spatiotemporal video transformer featured an innovative mechanism for decomposing spatiotemporal self-attention and self-reduction, enhancing robustness in deepfake detection.

Another study [56] focused on interframe motion across varying distances, proposing a dynamic difference Xi method for modelling spatiotemporal inconsistencies with precision. The study [57] achieved notable performance improvements by merging two feature extraction techniques: the YOLO facial detector and an enhanced HOG-based XceptionNet CNN. Study [58] investigated the role of action recognition methods in deepfake detection, comparing multiple networks and identifying the ResNet-based R3D model as delivering superior performance. Further, lightweight deep learning models have advanced in deepfake detection. In the study [59], they applied lightweight CNNs combined with sparse optical flow on facial regions, balancing high detection accuracy and reduced computational overhead. Referring to the study [47], the authors introduced a video deepfake detection approach utilizing a hybrid transformer-based architecture. Their approach employed EfficientNet-B0 for image feature extraction. These features were subsequently used to train two distinct Vision Transformer models: (1) Efficient ViT and (2) Convolutional Cross ViT. The latter model incorporated two branches: the S-branch, designed to handle images with smaller

patch sizes (7×7), and the L-branch, optimized for larger patch sizes (64×64), to provide a broader receptive field, a study evaluated on the DFDC dataset [60], demonstrated that the hybrid model outperformed other models tested in the study, combining EfficientNet-B0 with Convolutional Cross ViT. Additionally, cross-dataset testing on FaceForensics++ [11] yielded promising results.

In a study [61], they explored the application of transformer architecture in detecting deepfakes, proposing two innovative models: the Image Transformer and the Video Transformer. These models were trained using 3D facial features [62] and standard cropped face images, including 3D facial features aimed at aligning facial details more effectively and improving learning. The models could capture more relevant facial information by integrating these aligned features with traditional cropped face data. To utilize temporal information in videos, the researchers adapted the standard Vision Transformer (ViT) in a study [63] to process multiple sequential face frames. The model demonstrated incremental learning capabilities, effectively incorporating new data while retaining previously acquired knowledge. Extensive evaluations were performed using well-known deepfake detection benchmarks, including FaceForensics++ [11], DFDC [30], and Google DFD [64]. The results highlighted the model's strong performance across these datasets, emphasizing its effectiveness in detecting deepfakes.

These methods aim to identify temporal inconsistencies and classify videos at a high level without requiring additional aggregation steps. Some approaches target specific spatio-temporal artefacts commonly seen in deepfake videos, such as unusual lip movements [65] or irregular eye-blinking patterns [66]. However, they are somewhat limited as they do not explore other possible artefacts across different facial regions. Other techniques include optical flow analysis in videos [67] and examining the alignment between audio peaks and visual content [68]. A more recent method [69] introduces a two-stage process: the first stage employs self-supervised learning to extract temporal features from genuine videos, such as facial expressions and movements. The second stage leverages these features to train a forgery detection model to distinguish between real and fake videos. One notable approach is FTCN [70], which explicitly incorporates a Temporal Transformer Encoder to address temporal incoherence in videos. Despite their strengths, these methods often overlook critical aspects of the problem. For example, they usually consider only one subject in a video if multiple subjects are present, and do not consider variations in face-frame area ratios. One method in [52] tries to handle various subjects by considering each identity separately and classifying the whole video as fake in case manipulation in any of them is detected. In the work [71], the authors proposed the Multimodal Multi-scale Transformer model, which can process image patches with different sizes and thus detect local anomalies at various spatial levels. The M2TR model also developed frequency domain data and RGB information by the advanced cross-modality information fusion mechanism to improve its ability to identify forgery-related artefacts. Extensive experiments were conducted that established the efficacy of the M2TR model, which outperformed other deepfake detection models by notable margins.

The authors in [55] introduced an Interpretable Spatial-Temporal Video Transformer designed for deepfake detection. The model embeds a new decomposed spatio-temporal self-attention mechanism and self-subtraction that helps to identify spatial artefacts and temporal inconsistencies about forgery. Additionally, ISTVT facilitates visualizing the discriminative regions across both spatial and temporal dimensions with the help of a relevance propagation algorithm [55]. Extensive experiments on large-scale datasets confirmed that ISTVT performed very well in intra-dataset and inter-dataset deepfake detection, showing its effectiveness and robustness. Existing Deepfake detection models primarily target imperfections in the Deepfake generation process, such as irregularities in eye blinking [38,72]. However, with the rapid evolution of Deepfake technology, these imperfections are quickly addressed in newer models. For instance, a study [73] introduced a system capable of producing videos of talking heads with realistic facial expressions, including natural eye blinking. Similarly, the work done in [74], where the authors develop a model that

applies facial expressions from still images and can display emotions to the extent of even giving a blink-like motion illusion. All such detection methodologies are highly prone to defeat when enhanced Deepfake generation techniques are utilized.

This section of the literature review has underlined that the research community extensively deploys deep learning models, methods, and different techniques for developing effective and reliable systems of deepfake detection. However, scrutiny of these studies reveals that most often fail to provide similar performance in real scenarios or outside of their training distribution. To address these challenges, our study proposes a novel transformer-based framework for real-time deepfake detection through gaze and blink pattern analysis. The proposed framework uses the TimeSformer architecture, which has proven to perform exceptionally well in modelling spatiotemporal patterns in video data to tell apart manipulated and authentic media. Moreover, we embedded a hybrid Transformer-CNN model as a baseline for our comparison, thus enabling us to thoroughly investigate the trade-off between computational efficiency and detection accuracy.

3 Proposed Methodology

This section elaborates on the methods and techniques used and utilized in this novel study. This novel study aims to develop an innovative framework for detecting and analyzing real-time deepfake video with state-of-the-art transformer models, TimeSformer, and Transformer-CNN architecture. Our organized methodology covers the acquisition of widely used and key deepfake datasets, namely, FaceForensics [11], CelebDF-V2 [15], DFDC [30], and FakeAVCeleb [44]. These would offer a good source of authentic and manipulated videos, which will be important in evaluating the model's robustness across various manipulation techniques and demographics. Preprocessing mainly consists of extracting the region of interest (ROI), namely the eye region, through facial landmark detection, hence segmenting the relevant features concerning gaze and blink patterns. The extracted eye regions are then resized and normalized for consistent input dimensions across all frames. The noisy gaze trajectory and blink rate data are filtered using temporal smoothing techniques, such as Gaussian filtering, to make the temporal analysis cleaner and more accurate. At the core of the proposed methodology lies TimeSformer-a, a transformer-based architecture that segments frames into patches and applies a spatial attention mechanism to model different features around the eyes. The temporal attention of TimeSformer models the blinking intervals and gaze shifts, which are key indicators of authenticity in videos. This model directly deals with spatial and temporal patterns, making it ideal for finding anomalous behaviour in video sequences. A hybrid Transformer-CNN architecture integrates MesoNet4 and TimeSformer effectively; we introduced an intermediate fusion layer that consolidates the spatial features extracted by MesoNet4 with the temporal embeddings from TimeSformer. This approach optimizes spatial anomalies (e.g., inconsistent eye blinks) and temporal dynamics (e.g., gaze shifts) for forgery detection. Further, video augmentation techniques, such as frame cropping and ROI shifting, are used during training to enhance the robustness of the model.

The loss function used for training includes cross-entropy for classification tasks, with an auxiliary temporal consistency loss to address gaze and blink-related errors. Several metrics have been used for performance evaluation, including accuracy, precision, recall, and F1-score, together with specific gaze-related metrics like Gaze Direction Error (GDE) and Blink Detection Accuracy (BDA). Real-time performance metrics, such as inference time in FPS, are also essential in terms of the practical feasibility of the detection system. Further results are presented by comparing TimeSformer against the hybrid Transformer-CNN model and recent state-of-the-art works. This comparison includes the accuracy scores for all datasets and intra- and cross-dataset evaluations to determine the impact of single-dataset training. The methodology presents a holistic, effective, and robust approach toward eye movement- and temporal dynamics-based

detection of real-time deepfake videos, advancing state-of-the-art deep-fake detection techniques. The flow diagram of our proposed framework is shown in Fig. 1.

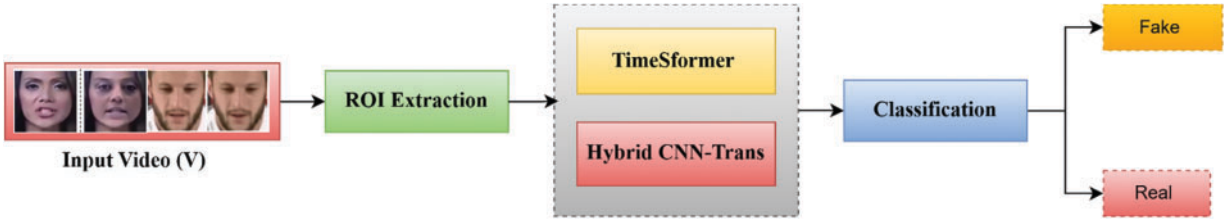

Figure 1: Flow diagram of proposed framework

3.1 Datasets

In this novel study, we have used four famous and widely employed benchmark datasets, namely FaceForensics [11], CelebDF-V2 [15], DFDC [30], and FakeAVCeleb [44]. These datasets are known to include real and synthetic videos on a wide scale and are generated using different deepfake creation techniques, offering various manipulation styles. We used these datasets to train and evaluate our proposed innovative framework to ensure robust performance and generalizability. Detailed insights into the datasets are discussed in the following sections concerning their novelty and individual value related to deepfake detection. These well-known datasets have also been utilized to present comprehensive training and testing, investigations, intra- and cross-dataset testing, and intra-comparison for the proposed novel framework. The specific attributes of each dataset are summarized in Table 1, providing insights into their role in enhancing the evaluation process.

Table 1: The real and fake images are used to train, validate, and test our models

Data distribution for the train/test data						
Dataset	**Train**		**Validation**		**Test**	
	Real	**Fake**	**Real**	**Fake**	**Real**	**Fake**
FaceForensics++ [11]	47,808	47,808	5360	5360	2000	2000
CelebDF-V2 [15]	50,000	50,000	10,000	10,000	1000	1000
DFDC [30]	50,000	50,000	10,000	10,000	2000	2000
FakeAVCeleb [44]	47,808	47,808	5360	5360	2000	2000

FaceForensics++ [11] is a prominent benchmark in deepfake detection, extensively utilized for evaluating the effectiveness of detection methodologies. Fig. 2 shows an example image from the FaceForensics++ dataset depicting several operations applied to the original video sequence.

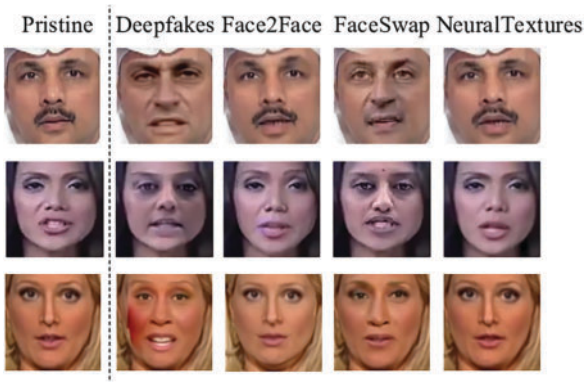

Figure 2: FaceForensics++ sample images

Fig. 3 presents an example image with manipulation methods of Deepfake synthesis from the CelebDF-V2 dataset.

Figure 3: CelebDF-V2 sample images

In Fig. 4, example image shows the deepfake generation process from the DFDC (Deepfake Detection Challenge) dataset. The image shows the diversity of the dataset concerning subjects' gender, ethnicity, and age.

Figure 4: DFDC sample images

FakeAVCeleb [44] dataset stands out as one of the most recently developed resources compared to other datasets utilized in this research. Fig. 5 presents an example image with manipulation methods of Deepfake media from the FakeAVCeleb dataset.

Figure 5: FakeAVCeleb sample images

3.2 Preprocessing

The preprocessing steps outlined in this study play a pivotal role in enabling the success of our novel deepfake detection framework. This step ensures data consistency and enhances the effectiveness of the deepfake detection framework. By isolating the eye region through precise Region of Interest (ROI) extraction leveraging the facial landmark detection, we separate key features for the eye region using (ROI), we focus directly on the most critical features related to gaze and blink patterns, which are fundamental for distinguishing authentic videos from manipulated ones. Further, normalization and resizing are regularized in the input dimensions for easy incorporation into the transformer-based model. Temporal smoothing cleans the data by removing noise and ensuring continuity, which is vital for dynamic patterns such as blinking and gaze shifting. Data augmentation improves robustness by simulating different scenarios and avoiding overfitting. These steps provide a good ground for spatial and temporal attention mechanisms in transformer-based architectures to enable efficient deepfake detection. The detailed mathematical formulations of these processes are represented below.

3.2.1 Region of Interest (ROI) Extraction

ROI extraction isolates the eye region, essential for analyzing gaze direction and blink patterns. Using Dlib's facial landmark detection, as shown in Fig. 6A, we identify the coordinates of facial landmarks $L = \{(x_i, y_i)\}_{i=1}^{n}$, where (x_i, y_i) represents the coordinates of the i-th landmark. From these coordinates, we define the bounding box B_{eye}, which surrounds the eye region.

To extract the eye region:

1. Define the bounding box B_{eye} around the eye landmarks:

$$B_{eye} = \left[\max_{i \in E} xi, \operatorname{man}_{i \in E} yi, \max_{i \in E} xi, \operatorname{man}_{i \in E} yi\right] \tag{1}$$

where E is the subset of indices corresponding to eye landmarks.

2. Crop the eye region from I using B_{eye}:

$$I_{eye} = I\left[B_{eye}\right] \tag{2}$$

This step produces consistent eye-region crops from all video frames, as depicted in Fig. 6B: ROI Extraction Process.

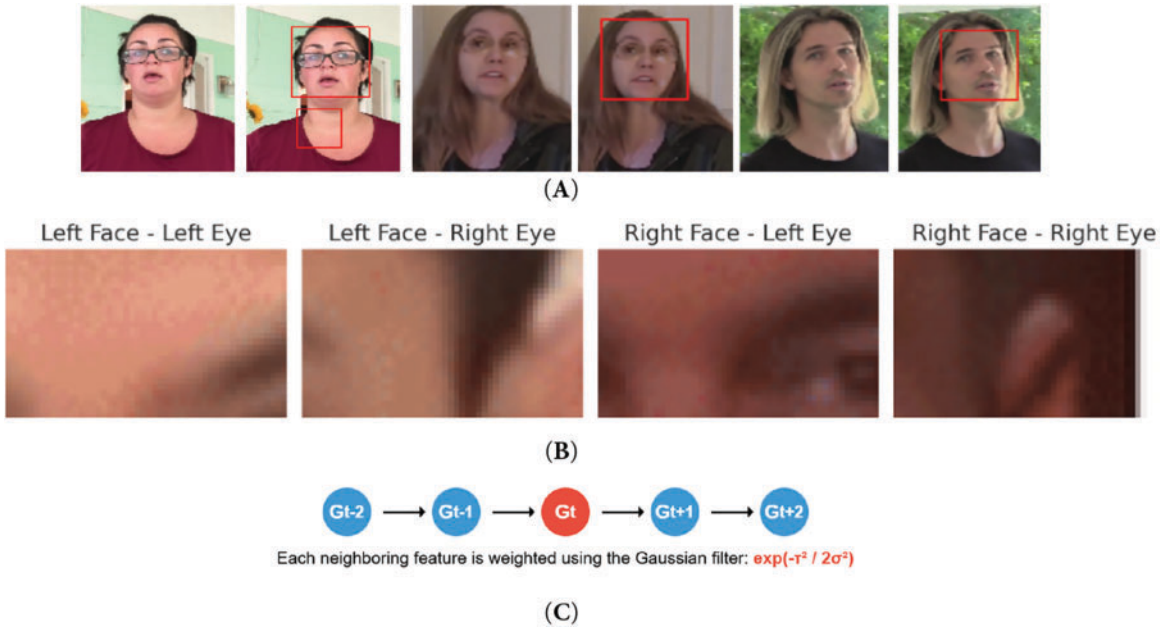

Figure 6: (A): ROI extraction focuses on facial landmark detection. **(B):** Cropped eye regions from two subjects for ROI extraction. **(C):** Illustration of temporal smoothing, each neighbouring feature is weighted using the Gaussian filter: $\exp(-\tau^2/2\sigma^2)$

3.2.2 Normalization and Resizing

The extracted eye regions I_{eye} are resized to a fixed dimension (H, W), to ensure uniformity across samples, where H and W are the target height and width. This is achieved using bicubic interpolation:

$$I'_{eye} = Resize(I'_{eye}, (H, W)) \tag{3}$$

Normalization is applied to scale pixel intensity values to the range $[0, 1]$, defined as:

$$I'_{eye}(i, j) = \frac{I'_{eye}(i, j) - \min(I'_{eye})}{\max(I'_{eye}) - \min(I'_{eye})} \tag{4}$$

This step standardizes input data for consistent model performance.

3.2.3 Temporal Smoothing

To reduce inconsistencies and noise in gaze and blink patterns, we apply a Gaussian filter for temporal smoothing, where the smoothed feature G_t at time t is computed as a weighted average of neighbouring gaze or blink features within a window size 2K + 1, as shown in Fig. 6C. The Gaussian filter is given by:

$$\widehat{G}_t = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{\tau=-K}^{K} G_{t+\tau} \exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{5}$$

where $G_{t+\tau}$ represents the neighbouring gaze/blink features, and σ controls the smoothing width.

Here:

- $\widehat{G_t}$: Smoothed feature at time t
- $G_{t+\tau}$: Neighboring features within a window size $2k + 1$
- σ: Standard deviation controlling the filter width

This process reduces inconsistencies in temporal features, enhancing downstream analysis.

3.2.4 Augmentation During Preprocessing

Augmentation is applied to improve the model's generalization of the extracted eye regions. This process includes:

1. ROI Shifting:

$$I''_{shifted}(i, j) = I''_{eye}(i + \Delta x, j + \Delta y) \tag{6}$$

where Δx and Δy are random offsets within a predefined range.

2. Cropping and Scaling: Randomly crop a sub-region and resize it to (H, W).

3. Brightness Adjustment:

$$I''_{aug} = \alpha I''_{eye} + \beta \tag{7}$$

where α and β control brightness and contrast adjustments.

3.3 TimeSformer Model

TimeSformer is a state-of-the-art model for video-based tasks, leveraging spatiotemporal attention mechanisms. This model is highly suitable for real-time detection of deepfake video content, as it can effectively model both spatial and temporal dependencies in video sequences. The proposed methodology leverages the strengths of this model in analyzing frame sequences directly and learning meaningful representations from eye-region features such as gaze shifts and blinking intervals. These are important for detecting manipulations in deepfake videos. The preprocessing pipeline establishes a critical foundation for our proposed TimeSformer model by seamlessly linking extracted features to the model's spatial and temporal analysis stages. Each video frame F_t undergoes ROI extraction, isolating the eye region, represented as F_t^{eye}, through the function f_{ROI}. This F_t^{eye} ensures that only the most relevant features related to gaze and blink dynamics are processed. The output F_t^{eye} is then subjected to normalization, resizing, and partitioning into spatial patches, denoted as $\mathrm{F}_t^{\mathrm{patches}e} = \{P_{1,1}, P_{1,2,} \ldots, P_{m,n,}\}$, to meet the input requirements of TimeSformer's attention mechanisms. The preprocessing pipeline not only provides high-quality and consistent data for model input but also introduces robust augmentation strategies that will make F_t^{eye} more resilient to spatial distortions, such as ROI shifting and random cropping. These thoughtfully designed preprocessing steps bridge the gap between raw video data and the sophisticated spatial-temporal capabilities of the TimeSformer model by assuring robust feature extraction, hence improving our model performance.

Mathematical Formulation

Let the input video V be composed of T frames denoted as:

$$V = \{F1, F2, \ldots, FT\} \tag{8}$$

Each frame F_t undergoes ROI extraction to isolate the eye region using a function f_{ROI}:

$$\mathrm{F}_t^{eye} = f_{ROI}(F_t) \tag{9}$$

Next, the extracted regions are divided into spatial patches P, where $P_{i,j}$ represents the (i, j) – th patch in a frame:

$$\mathrm{F}_t^{\text{patchese}} = \{P_{1,1}, P_{1,2}, \ldots, P_{m,n}\} \tag{10}$$

For each patch $P_{i,j}$, positional embeddings e_{pos} are added to encode location information:

$$Z_0 = P_{i,j} + e_{\text{pos}} \tag{11}$$

The spatial attention mechanism processes these embeddings using self-attention to compute relationships between patches within the frame:

$$Z_{\text{spatial}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V}, \tag{12}$$

where Q, K, and V are the query, key, and value matrices extracted from the embedded image patches.

For temporal modelling, the sequence of spatial embeddings across frames is processed to capture temporal dynamics:

$$Z_{\text{temporal}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\mathrm{V}, \tag{13}$$

where the inputs are now the spatially processed embeddings from consecutive frames.

The final feature representation for the video is obtained by concatenating or aggregating the spatial and temporal embeddings:

$$Z_{\text{final}} = \text{Concat}\left(Z_{\text{spatial}}, Z_{\text{temporal}}\right). \tag{14}$$

This representation is fed into a classification head for distinguishing real vs. deepfake videos.

TimeSformer model employs a cutting-edge approach to video analysis by segmenting video frames into smaller non-overlapping patches, treating each patch as an input token analogous to words in natural language processing. By including architecture leveraging two dimensions of self-attention, this first incorporates spatial attention for intra-frame feature relationships by considering layout and interactions between the detected regions of interest, or simply eye regions. Second, temporal attention serves for inter-frame dependency modelling, which means the dynamics of shifts in gaze and blinking subsequently take variations concerning time. This dual attention mechanism enables TimeSformer to capture both the spatial and temporal consistencies, which is crucial to understanding whether a video is authentic or manipulated. Table 2 provides a clear overview of all the parameters and the overall configuration settings of the TimeSformer model. The complete working process of the proposed detection system is outlined in Algorithm 1.

Table 2: Parametric table for TimeSformer model

Parameter	Description	Value
Input size	Resolution of extracted eye regions	128 × 128
Patch size	Size of each spatial patch	16 × 16
Sequence length	Number of video frames	32
Embedding dimension	Dimensionality of embeddings	768
Number of layers	Transformer layers	12
Attention heads	Multi-head attention count	8
Dropout rate	Regularization dropout	0.1
Optimizer	Training optimizer	AdamW
Learning rate	Initial learning rate	3×10^{-4}
Epochs	Number of training iterations	50

Algorithm 1: Processing eye features using TimeSformer

Input: Video $V = \{F1, F2, \ldots, FT\}$

Output: Classification of authenticity $y \in \{\boldsymbol{real}, \boldsymbol{fake}\}$.

1. **Initialization:**
   - Instantiate ROI extractor f_{ROI}
   - Initialize TimeSformer model $\boldsymbol{TimeSformer}$
2. **Preprocessing:**
   - For each frame $F \in V$:
   - Extract eye region: $\mathrm{F}_t^{eye} \leftarrow f_{ROI}(F_t)$
   - Normalize and resize F_t^{eye}
3. **Feature Extraction:**
   - Divide F_t^{eye} into spatial patches $\mathrm{F}_t^{\mathrm{patches}e}$.
   - Add positional embeddings to each patch.
4. **Attention Mechanisms:**
   - Apply spatial attention to patches within each frame.
   - Apply temporal attention across frames.
5. **Classification:**
   - Aggregate final embeddings Z_{final}
   - Pass through the classification head to predict y.

Return: Predicted label y.

The TimeSformer model has several distinct advantages, making it a robust choice for deepfake detection. The integrated approach to spatiotemporal dynamics removes the need to separately process spatial and temporal information, thus allowing it to detect subtle manipulations in videos effectively. TimeSformer model attains high proficiency by dividing video frames into smaller patches, reducing computational demands while retaining essential details and enhancing scalability for large datasets. Attention mechanisms enhance robustness by focusing on the foremost region, like the eyes, which show noticeable variations in gaze and blinking due to irregularities. Fig. 7 represents the flexible TimeSformer model architecture, which thus generalizes well into varied datasets with manipulation techniques for applicability in wide Deepfake Detection Scenarios.

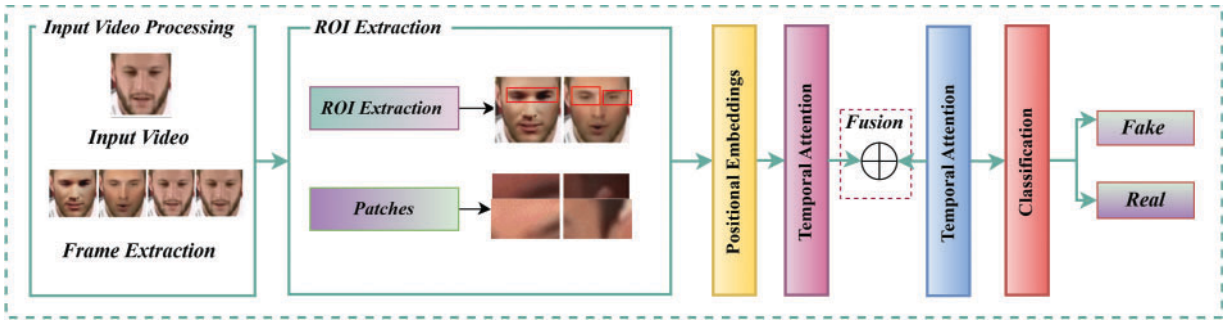

Figure 7: TimeSformer model architecture

3.4 Hybrid Transformer-CNN

To evaluate the efficiency of our primary model, TimeSformer, we introduce an alternative approach: a Hybrid Transformer-CNN model. The idea is to combine a CNN's spatial feature extraction capability with the temporal attention mechanism that transformers can offer, providing a complementary architecture for deepfake detection. By integrating MesoNet4 for spatial analysis and a lightweight transformer module for temporal modelling, this alternative model serves as a benchmark, highlighting the advantages of the fully transformer-based TimeSformer architecture. The Hybrid Transformer-CNN model processes input video frames through spatial feature extraction and temporal modelling. First, MesoNet4 isolates and analyzes gaze and blink features from eye regions. These extracted spatial features are later fed into a transformer module to model temporal dependencies across frames for detecting anomalies in video sequences. Fig. 8 represents the hybrid transformer-CNN model architecture.

Mathematical Formulation

Let the video sequence $V = \{F_1, F_2, \ldots, F_T\}$ consist of T frames. Each frame F_t undergoes ROI extraction to isolate the eye region, denoted as F_t^{eye}, using a function f_{ROI}:

$$F_t^{eye} = f_{ROI}(F_t) \tag{15}$$

The CNN component processes the extracted ROI to generate spatial features S_t:

$$S_t = CNN\left(F_t^{eye}\right) \tag{16}$$

The transformer module takes the sequence of spatial features $\{S_1, S_2, \ldots, S_T\}$ as input, applying temporal attention to model inter-frame dependencies:

$$H_t = \text{Transformer}\ (S_t) \tag{17}$$

The final representation H_t is classified into real or fake using a fully connected classification head:

$$y = \text{Classifier}\ (H_t) \tag{18}$$

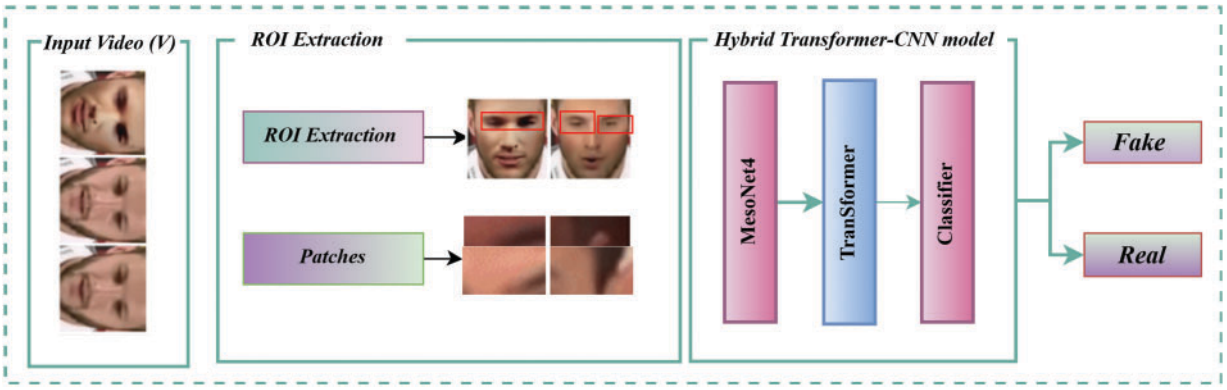

Figure 8: Hybrid transformer-CNN model architecture

The Hybrid Transformer-CNN model demonstrates notable advantages, including enhanced efficiency through MesoNet4. The CNN backbone in our hybrid model uses the MesoNet4 architecture, which employs a kernel size of 3 × 3 with a stride of 1 for convolutional operations. These choices help capture fine-grained features from the eye region while retaining spatial resolution. These hyperparameters are listed in Table 3. Hybrid Transformer-CNN reduces computational complexity compared to full-transformer architectures. Its modular design, leveraging independent spatial and temporal processing, ensures adaptability across various video qualities and resolutions. To extract eye-blink features, the step-by-step procedure is detailed in Algorithm 2. Furthermore, its robust feature fusion mechanisms strengthen the model's resilience to subtle temporal inconsistencies, enabling more accurate detection of deepfake manipulations. Table 3 shows all the parameters and the detailed Hybrid Transformer-CNN Model parameter settings.

Table 3: Parametric table for hybrid transformer-CNN model

Parameter	Description	Value
Input size	Resolution of extracted eye regions	128 × 128
Embedding dimension	Dimensionality of embeddings	768 (aligned with TimeSformer)
CNN Backbone	Spatial feature extractor	MesoNet4
Kernel size	Size of convolutional kernels in CNN layers	3 × 3
Stride	Stride used in CNN layers	1
Transformer layers	Temporal attention layers	4
Attention heads	Multi-head attention count	4
Dropout rate	Regularization dropout	0.2
Optimizer	Training optimizer	Adam
Learning rate	Initial learning rate	1×10^{-3}
Epochs	Number of training iterations	50
Batch size	Training batch size	32

Algorithm 2: Hybrid Transformer-CNN for eye feature processing

Input: Video $V = \{F_1, F_2, \ldots, F_T\}$
Output: Classification $y \in \{\text{real, fake}\}$

(Continued)

Algorithm 2 (continued)

1. **Initialization:**
   - Instantiate ROI extractor f_{ROI}.
   - nitialize MesoNet4 for spatial feature extraction.
   - Initialize lightweight Transformer for temporal modelling.
2. **Preprocessing:**
   - For each frame $F_t \in V$:
   - Extract eye region: $F_t^{cyc} \leftarrow f_{ROI}(F_t)$.
   - Normalize and resize F_t^{cyc}.
3. **Feature Extraction:**
   - Pass F_t^{eye} through CNN to extract spatial features: $S_t \leftarrow CNN\left(F_t^{eyc}\right)$.
4. **Temporal Attention:**
   - Process spatial features through the Transformer to capture temporal dependencies: $H_t \leftarrow$ Trans former (S_t).
5. **Classification:**
   - Pass H_t through a fully connected layer to predict y.

Return: Predicted label y.

The Hybrid Transformer-CNN model offers a robust baseline for deepfake detection by combining the lightweight MesoNet4 architecture for spatial feature extraction with TimeSformer's temporal attention mechanism, balancing computational efficiency and detection capability. Its modular design ensures adaptability to varying video qualities and resolutions, while the fusion of spatial and temporal features via intermediate-layer integration enhances its resilience to subtle manipulation artefacts. However, as demonstrated in the comparative results, the TimeSformer model exhibits superior performance, surpassing the Hybrid Transformer-CNN in critical metrics such as accuracy, precision, recall, F1-score, GDE, BDA, and FPS. This is particularly evident in scenarios involving complex temporal inconsistencies, where TimeSformer's integrated spatiotemporal attention mechanisms provide a more comprehensive analysis of video patterns. These results underscore the advanced capabilities of TimeSformer, making it the preferred model for detecting sophisticated deepfake manipulations.

3.5 Evaluation Metrics

This study assessed the efficiency and robustness of our proposed novel framework using advanced real-time performance metrics, particularly domain-specific metrics, such as gaze and blink detection. Each performance metric is discussed below.

3.5.1 Accuracy

The accuracy metric is the ratio of actual and predicted real and fake instances over the total cases. This metric gives a general overview of the model's performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where:

- *TP*: True Positives (correctly classified fake videos),
- *TN*: True Negatives (correctly classified real videos),
- *FP*: False Positives (real videos misclassified as fake),

- *FN*: False Negatives (fake videos misclassified as real).

3.5.2 Precision

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. This is an essential metric in settings where false alarms (misclassifying real videos as fake) must be minimized.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{20}$$

3.5.3 Recall (Sensitivity)

Recall measures the proportion of actual positive instances correctly identified by the model. High recall ensures that most deepfake videos are detected, which is vital for security applications.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{21}$$

3.5.4 F1-Score

The F1 score provides a balanced measure; it considers the model's efficiency in identifying true positives and its precision.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

3.5.5 Gaze Direction Error (GDE)

The GDE (Gaze Direction Error) measures the difference between the predicted and actual gaze vectors. GDE quantifies how much the model captures gaze inconsistencies, one of the most prominent cues for deepfake manipulations.

$$\text{GDE} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{g}_i^{\text{true}} - \mathbf{g}_i^{\text{pred}} \right\| \tag{23}$$

where:

- $\mathbf{g}_i^{\text{true}}$: Ground truth gaze vector for frame *ii*,
- $\mathbf{g}_i^{\text{pred}}$: Predicted gaze vector for frame *ii*,
- N: Total number of frames.

3.5.6 Blink Detection Accuracy (BDA)

The BDA (Blink Detection Accuracy) metric is related to the correctness of the blink prediction, which is a significant feature in deepfake detection. BDA reflects the model's sensitivity to temporal features such as blinking patterns.

$$\text{BDA} = \frac{\text{Number of Correct Blink Detections}}{\text{Total Blinks (Ground Truth)}} \tag{24}$$

3.5.7 Inference Time (Frames Per Second, FPS)

The metric (Frames Per Second, FPS) is used to interpret real-time performance measures of a model in terms of (FPS) frames processed per second. A high FPS value is desirable for arranging deepfake detection systems in real-time scenarios.

$$\text{FPS} = \frac{\text{Number of Frames Processed}}{\text{Total Processing Time}} \tag{25}$$

3.5.8 Confusion Matrix

The CM metric is essential in model performance to break for all classification classes in detail as shown in Table 4. CM matrix granularly details errors and corrects predictions to find model behaviour insights.

Table 4: Presents a generic structure for the confusion matrix used in this study

Actual/Predicted	Fake	Real
Fake	TP	FNFN
Real	FP	TNTN

The evaluation results emphasize the superior performance of the TimeSformer model, which consistently outperforms the Hybrid Transformer-CNN across all metrics due to its robust spatiotemporal attention mechanisms.

4 Results and Discussion

The result section highlights an in-depth evaluation assessment and analysis of the performance of our novel framework for real-time deepfake video detection based on analyzing gaze and blink patterns using the TimeSfsssormer transformer model and hybrid Transformer-CNN architecture. Our novel study used diverse datasets, including FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb, to assess our state-of-the-art work's robustness and efficiency. Further, to validate the study, we compare our TimeSformer transformer-based model with our hybrid transformer-CNN model. To analyze the credibility and effectiveness of our proposed framework, comparisons were made with existing state-of-the-art methods. The comparison was carried out to determine the model's capability in distinguishing between real and manipulated media under diverse circumstances. Table 5 and Fig. 9 show that TimeSformer model performance is evaluated based on accuracy, precision, recall, F1-score, GDE, BDA, and FPS score metrics. Fig. 10 also presents the respective confusion matrices to provide an additional glimpse into the classification performance. Similarly, Table 6 and Fig. 11 summarize the evaluation outcomes of the hybrid Transformer-CNN model.

Table 5: Performance metrics of TimeSformer across datasets. The most significant improvements in the metrics are highlighted in bold. Notably, ssTimeSformer shows a 4.7% improvement in accuracy on the FaceForensics dataset and a 2.3% increase in F1-score on FakeAVCeleb, outperforming the Hybrid Transformer-CNN model

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	**97.5**	**97.8**	**97.2**	**97.4**	**1.5°**	**90.8**	**30**
CelebDF-V2	96.3	96.5	96.1	96.3	1.8°	89.3	28

(Continued)

Table 5 (continued)

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
DFDC	95.8	95.9	95.6	95.7	2.1°	88.5	27
FakeAVCeleb	97.1	97.3	97.0	97.2	1.6°	91.0	29

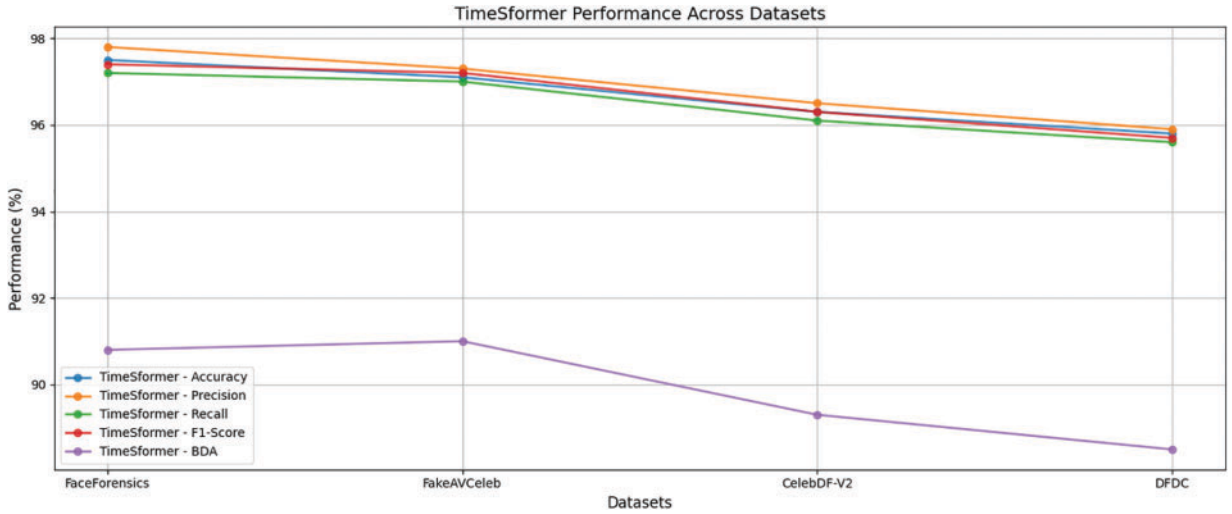

Figure 9: Performance metrics of TimeSformer across datasets. TimeSformer demonstrates significant improvements across multiple datasets, including a 4.7% increase in accuracy on FaceForensics (97.5% vs. 92.8%) and higher precision and recall on CelebDF-V2 (+4.4% and +5.1%, respectively). It also outperforms with a lower GDE (1.5° vs. 3.0° on FaceForensics), 5.6% higher BDA on FakeAVCeleb (91.0% vs. 86.3%), and faster FPS (30 FPS vs. 24 FPS on FaceForensics), underscoring its enhanced accuracy, efficiency, and ability to detect subtle deepfake manipulations

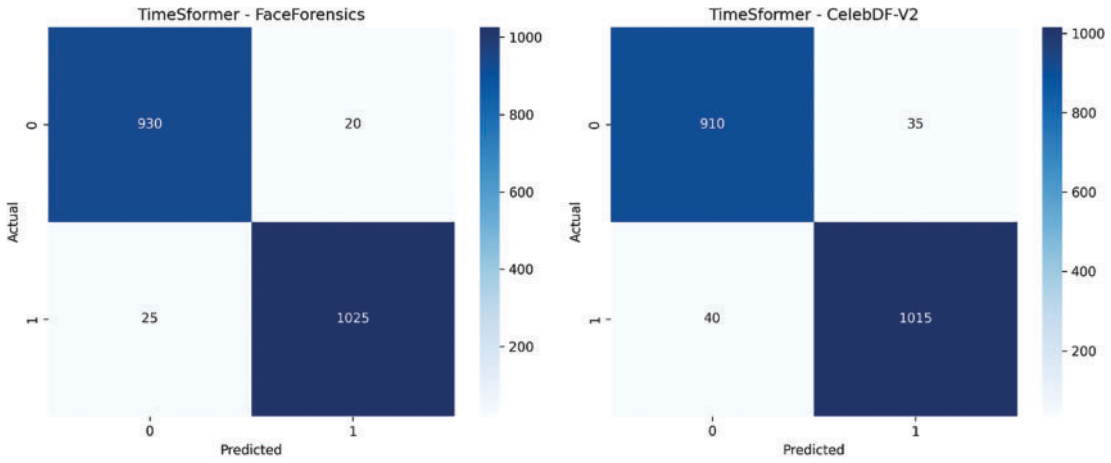

Figure 10: (Continued)

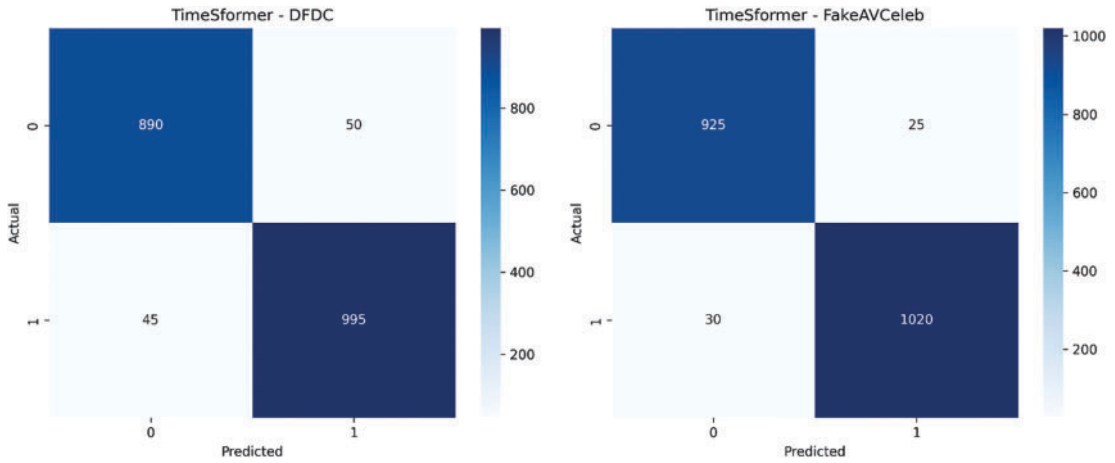

Figure 10: Confusion matrices for TimeSformer across four datasets

Table 6: Hybrid transformer-CNN performance across datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	92.8	93.5	92.2	92.9	3.0°	85.2	24
CelebDF-V2	91.5	92.1	91.0	91.8	3.4°	83.7	23
DFDC	90.9	91.7	90.5	91.2	3.6°	82.9	21
FakeAVCeleb	93.2	94.0	93.0	93.5	3.2°	86.3	22

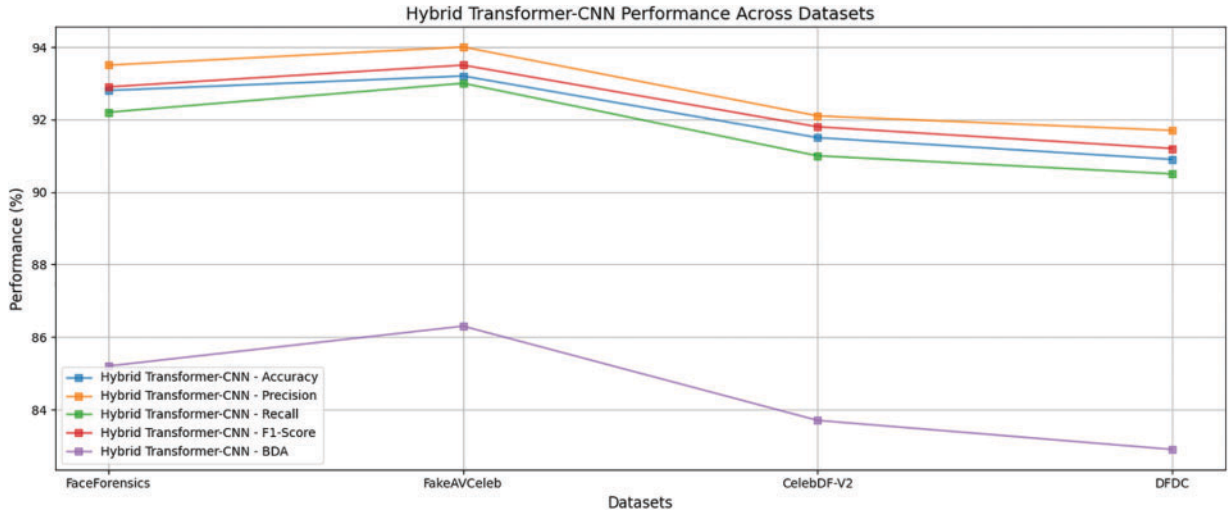

Figure 11: Hybrid transformer-CNN performance across datasets

In contrast, Fig. 12 presents its confusion matrices to facilitate a detailed analysis of its predictive behavior. The selected datasets contain various challenges like manipulation diversity, subject variability, and environmental conditions. Together, these evaluations comprehensively view the suggested models' strengths and weaknesses.

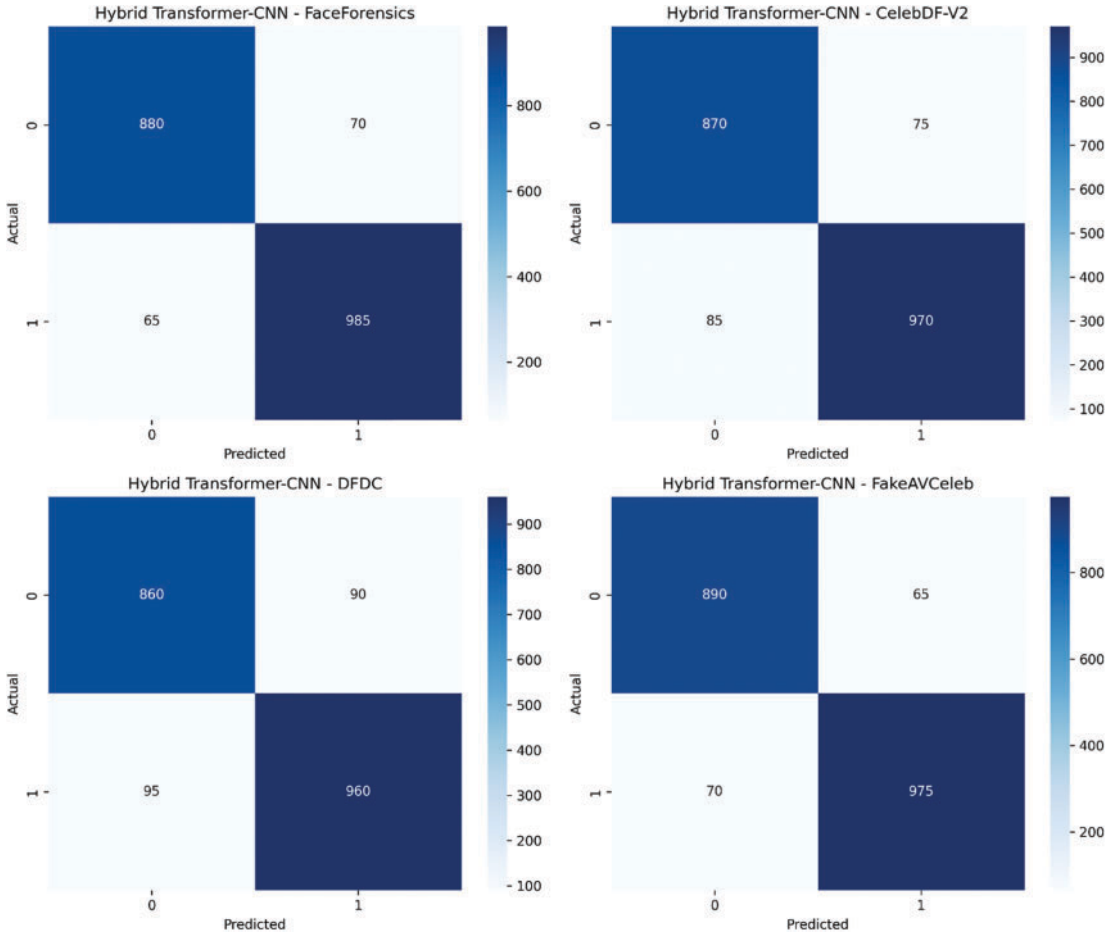

Figure 12: Confusion matrices for hybrid transformer-CNN across four datasets

For the FaceForensics dataset, the TimeSformer transformer-based model outperforms an accuracy of 97.5, precision of 97.8, detection rate of 97.2, F1-score of 97.4, GDE 1.5°, BDA 90.8, and FBS 30. The corresponding confusion matrices [930, 20, 25, 1025] provided a detailed breakdown of true positives, true negatives, false positives, and false negatives. Fig. 10 shows the confusion matrix result of the proposed TimeSformer transformer-based model on FaceForensics for real-time Deepfake video detection.

For the CelebDF-V2 dataset, the TimeSformer transformer-based model outperforms an accuracy of 96.3, precision of 96.5, detection rate of 96.1, F1-score of 96.3, GDE 1.8°, BDA 89.3, and FBS 28. The confusion matrices [910, 35, 40, 1015] give an exhaustive view of the model's accuracy in classification through true positives, true negatives, false positives, and false negatives. Fig. 10 demonstrates the confusion matrix result of the proposed TimeSformer-based model on the CelebDF-V2 dataset, with its real-time deepfake video detection performance.

For the DFDC dataset, the TimeSformer transformer-based model outperforms an accuracy of 95.8, precision of 95.9, detection rate of 95.6, F1-score of 95.7, GDE 2.1°, BDA 88.5, and FBS 27. The confusion matrices [890, 50, 45, 995] show a detailed breakdown of the classification result in true positives, true negatives, false positives, and false negatives. Fig. 10 presents the confusion matrix of the proposed TimeSformer-based model evaluated on the DFDC dataset, showing its performance in real-time deepfake video detection. For the FakeAVCeleb dataset, the TimeSformer transformer-based model outperforms an

accuracy of 97.1, precision of 97.3, detection rate of 97.0, F1-score of 97.2, GDE 1.6°, BDA 91.8, and FBS 29. The confusion matrix [925, 25, 30, 1020] clearly breaks down the model's classification performance in terms of true positives, true negatives, false positives, and false negatives. Fig. 10 illustrates the result of the proposed TimeSformer-based model on the FakeAVCeleb dataset, showing its effectiveness in real-time deepfake video detection.

For the FaceForensics dataset, the Hybrid Transformer-CNN model achieves an accuracy of 92.8%, a precision of 93.5%, a detection rate of 92.2%, an F1-score of 92.9%, GDE 3.0°, BDA 85.2, and FBS 24. Confusion matrix [880, 70, 65, 985] is the fine-grained classification outcome perspective that displays true positives, true negatives, false positives, and false negatives. Fig. 12 shows the confusion matrix of the proposed Hybrid Transformer-CNN model tested on the FaceForensics dataset, revealing the model's real-time deepfake video detection performance.

For the CelebDF-V2 dataset, the Hybrid Transformer-CNN model achieves an accuracy of 91.5, a precision of 92.1, a detection rate of 91.0, an F1-score of 91.8, GDE 3.4°, BDA 83.7, and FBS 23. The confusion matrix [870, 75, 85, 970] illustrates the total failure of classification results, such as true positives, true negatives, false positives, and false negatives. Fig. 12 illustrates the performance of the proposed Hybrid Transformer-CNN model on the CelebDF-V2 dataset, noting its effectiveness in real-time deepfake video detection.

For the DFDC dataset, the Hybrid Transformer-CNN model achieves an accuracy of 90.9, precision of 91.7, detection rate of 90.5, F1-score of 91.2, GDE 3.6°, BDA 82.9, and FBS 21. The confusion matrix [860, 90, 95, 960] offers a complete description of classification performance, including true positives, true negatives, false positives, and false negatives. Fig. 12 illustrates the hybrid Transformer-CNN model's output of the confusion matrix in the case of the DFDC dataset to emphasize its effectiveness in real-time detection of deepfake videos.

For the FakeAVCeleb dataset, the Hybrid Transformer-CNN model achieves an accuracy of 93.2, precision of 94.0, detection rate of 93.0, F1-score of 93.5, GDE 3.2°, BDA 86.3, and FBS 22. The confusion matrix [890, 65, 70, 975] gives a detailed description of classification accuracy, which identifies true positives, true negatives, false positives, and false negatives. Fig. 12 shows the confusion matrix of the proposed Hybrid Transformer-CNN model evaluated on the FakeAVCeleb dataset, showing its real-time deepfake video detection capability.

We compare the TimeSformer model evaluation result across datasets with the Hybrid Transformer-CNN model performance across datasets in Table 7 and Fig. 13. The TimeSformer model showed a promising result against the Transformer-CNN model performance.

Table 7: Comparison of TimeSformer vs. hybrid transformer-CNN across datasets

Metric	Dataset	TimeSformer	Hybrid Transformer-CNN
Accuracy (%)	FaceForensics	97.5	92.8
	CelebDF-V2	96.3	91.5
	DFDC	95.8	90.9
	FakeAVCeleb	97.1	93.2
Precision (%)	FaceForensics	97.8	93.5
	CelebDF-V2	96.5	92.1
	DFDC	95.9	91.7

(Continued)

Table 7 (continued)

Metric	Dataset	TimeSformer	Hybrid Transformer-CNN
Recall (%)	FakeAVCeleb	97.3	94.0
	FaceForensics	97.2	92.2
	CelebDF-V2	96.1	91.0
	DFDC	95.6	90.5
F1-Score (%)	FakeAVCeleb	97.0	93.0
	FaceForensics	97.4	92.9
	CelebDF-V2	96.3	91.8
	DFDC	95.7	91.2
GDE (°)	FakeAVCeleb	97.2	93.5
	FaceForensics	1.5°	3.0°
	CelebDF-V2	1.8°	3.4°
	DFDC	2.1°	3.6°
BDA (%)	FakeAVCeleb	1.6°	3.2°
	FaceForensics	90.8	85.2
	CelebDF-V2	89.3	83.7
	DFDC	88.5	82.9
FPS	FakeAVCeleb	91.0	86.3
	FaceForensics	30	24
	CelebDF-V2	28	23
	DFDC	27	21
	FakeAVCeleb	29	22

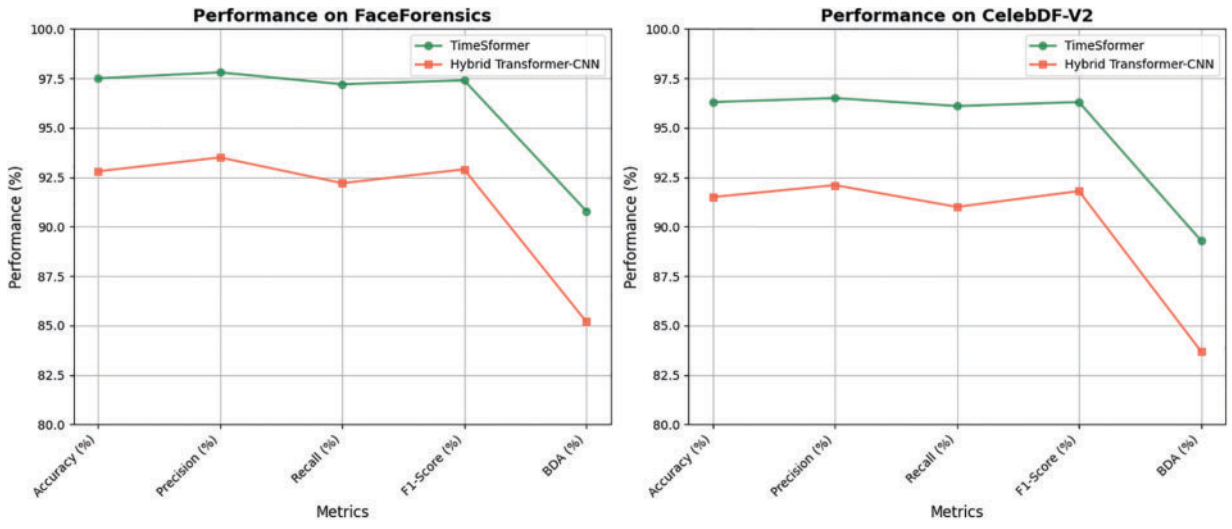

Figure 13: (Continued)

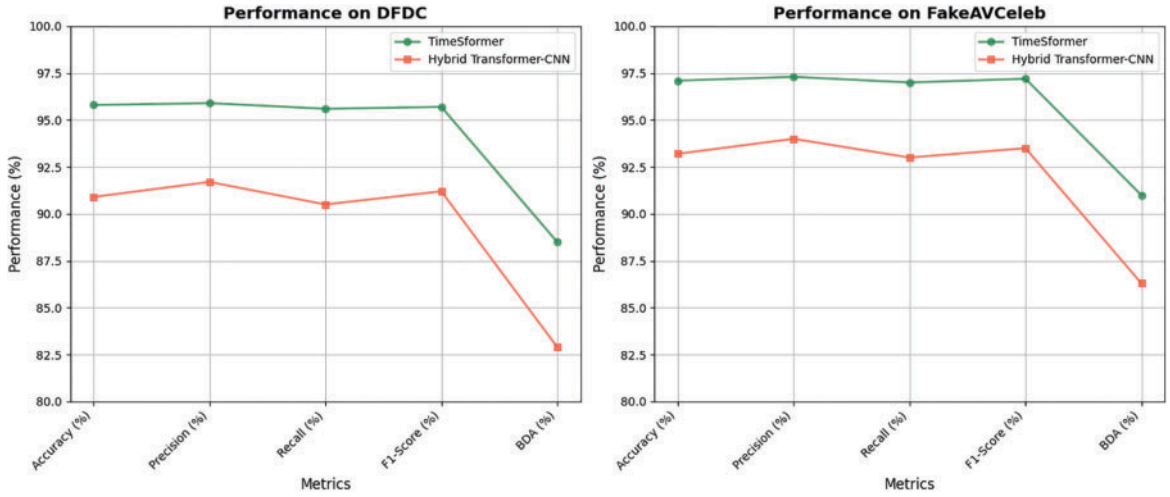

Figure 13: Comparison of TimeSformer and hybrid transformer-CNN across datasets

The comparative evaluation of the TimeSformer and Hybrid Transformer-CNN models across datasets demonstrates the clear superiority of the TimeSformer in terms of accuracy, precision, recall, and F1-score, as evidenced in Table 7 and the accompanying confusion matrices in Figs. 10 and 12. Specifically, TimeSformer consistently achieved over 95% accuracy across all datasets, with powerful performances on FaceForensics (97.5%) and FakeAVCeleb (97.1%), while maintaining lower GDE values (e.g., 1.5° for FaceForensics) and higher BDA scores. In contrast, the Hybrid Transformer-CNN displayed moderate metrics, with accuracies ranging from 90.9% (DFDC) to 93.2% (FakeAVCeleb) and higher GDE values (e.g., 3.6° for DFDC), indicative of reduced precision in detecting deepfake manipulations. The comparative confusion matrices reinforce these findings, showcasing higher true favourable and true negative rates for TimeSformer, reflecting its capability to minimize false negatives and false positives. Overall, the technical evaluation underscores the advanced temporal and spatial modelling of TimeSformer, making it a robust choice for real-time deepfake detection. At the same time, the Hybrid Transformer-CNN, though competitive, falls short in addressing finer manipulation details.

Tables 8–11 and Figs. 14–17 show the Intra and cross-dataset evaluation of the TimeSformer model on the FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb datasets.

Table 8: Intra and cross-dataset evaluation of Timesformer when trained on faceforensics dataset

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	97.5	97.8	97.2	97.4	1.5°	90.8	30
CelebDF-V2	95.6	95.8	95.2	95.5	1.7°	88.5	28
DFDC	94.8	95.1	94.3	94.7	1.8°	87.2	27
FakeAVCeleb	95.2	95.5	94.7	95.1	1.6°	88.0	29

Table 9: Intra and cross-dataset evaluation of TimeSformer when trained on CelebDF-V2 dataset

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	94.8	95.0	94.3	94.6	2.1°	87.5	30
CelebDF-V2	96.3	96.5	96.1	96.3	1.8°	89.3	28
DFDC	94.1	94.3	93.5	93.9	2.3°	85.2	26
FakeAVCeleb	94.2	94.5	93.8	94.0	2.0°	86.1	27

Table 10: Intra and cross-dataset evaluation of TimeSformer when trained on DFDC dataset

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	94.8	95.1	94.3	94.6	2.3°	87.1	28
CelebDF-V2	94.1	94.3	93.8	94.0	2.2°	86.4	27
DFDC	95.8	95.9	95.6	95.7	2.1°	88.5	27
FakeAVCeleb	94.3	94.5	94.0	94.2	2.4°	86.7	28

Table 11: Intra and cross dataset evaluation of TimeSformer when trained on the FakeAVCeleb dataset

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	GDE (°)	BDA (%)	FPS
FaceForensics	94.7	94.9	94.2	94.6	1.5°	87.5	30
CelebDF-V2	94.0	94.3	93.6	94.0	1.4°	86.8	29
DFDC	94.1	94.4	93.8	94.1	1.4°	86.5	28
FakeAVCeleb	97.1	97.3	97.0	97.2	1.6°	91.0	29

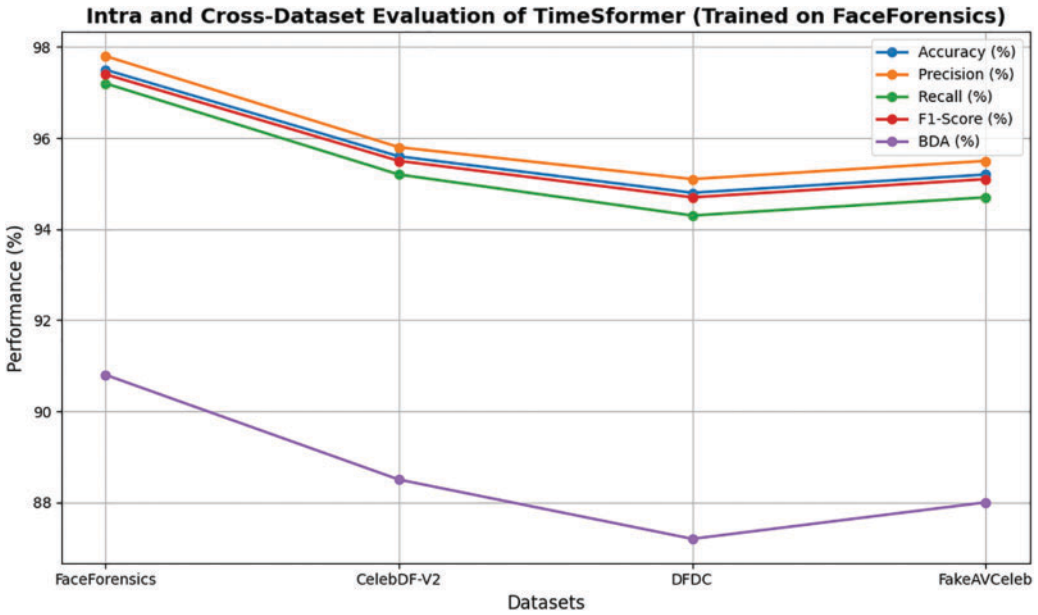

Figure 14: Intra and cross-dataset evaluation of TimeSformer (Trained on FaceForensics)

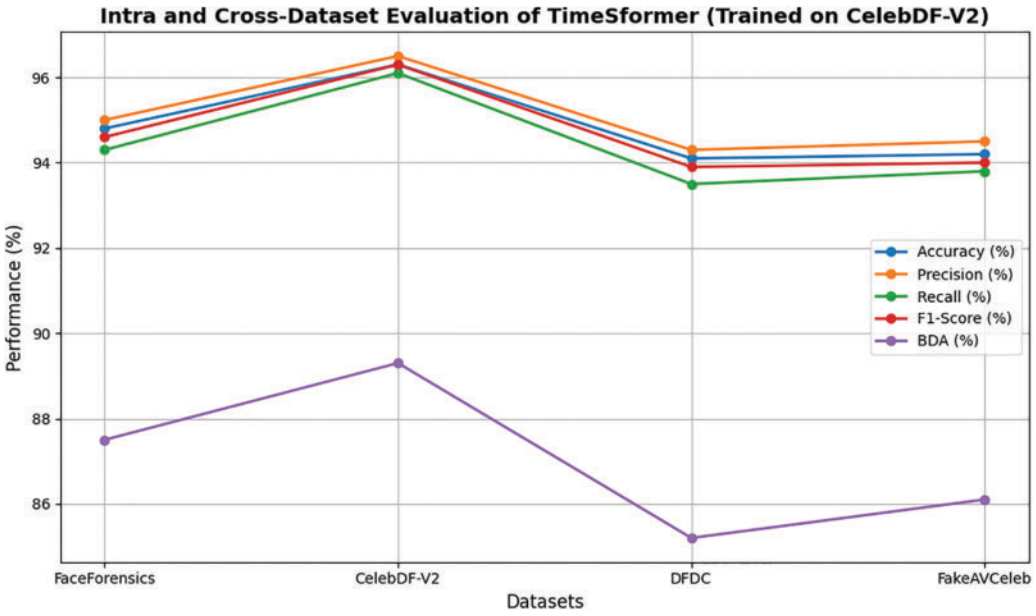

Figure 15: Intra and cross-dataset evaluation of TimeSformer (Trained on CelebDF-V2)

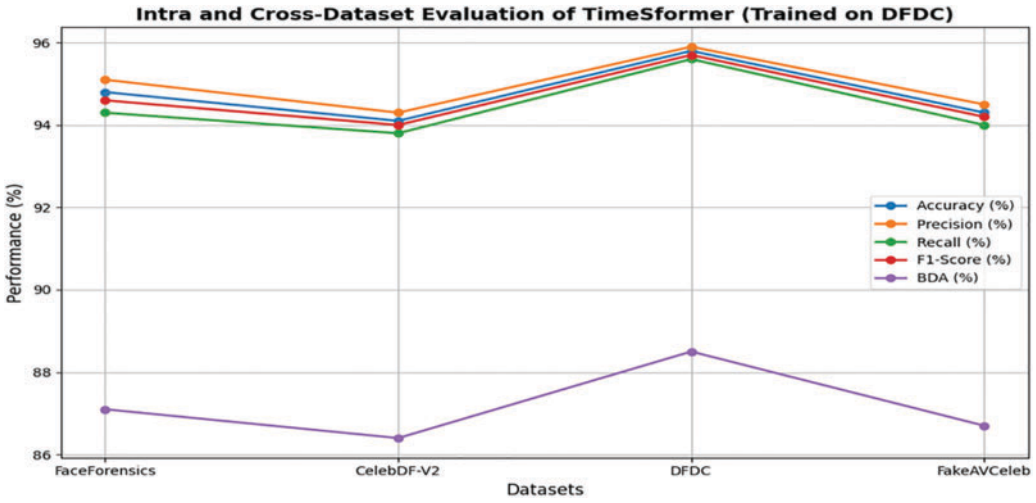

Figure 16: Intra and cross-dataset evaluation of TimeSformer (Trained on DFDC)

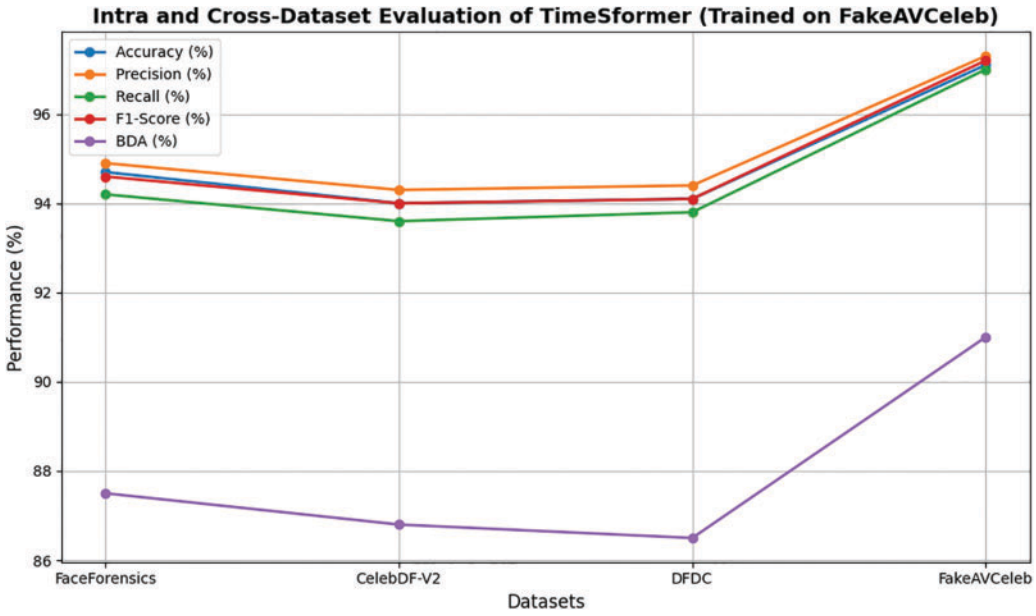

Figure 17: Intra and cross-dataset evaluation of TimeSformer (Trained on FakeAVCeleb)

The TimeSformer model outperforms in intra- and cross-dataset evaluation with the highest accuracy of 97.5 on the FaceForensics dataset. Tables 8–11 show the intra- and cross-dataset evaluation of the TimeSformer model on the FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb datasets. The evaluation of our novel framework, employing TimeSformer and Hybrid Transformer-CNN models, is rigorously assessed using widely used performance metrics and compared against state-of-the-art methods across multiple datasets in Table 12. A comparative analysis in Table 12 highlights our proposed approaches' superior performance against previous Deepfake video detection studies. Notably, the TimeSformer model demonstrates exceptional accuracy, achieving (97.5%) on the FaceForensics dataset with high precision (97.8%) and recall (97.2%), resulting in an F1-score of 97.4%. These results emphasize the model's ability to accurately classify manipulated and genuine videos, with minimal false positives and false negatives, as further evidenced by the confusion matrices. Likewise, though slightly inferior in terms of performance to TimeSformer, the Hybrid Transformer-CNN approach still provides robust results. For instance, the FakeAVCeleb dataset attained an accuracy of 93.2%, a precision of 94.0%, a recall of 93.0%, and an F1-score of 93.5%, proving its competitive efficiency in real-time applications. Execution times per frame were also considered, and the results are summarized in Table 7. These times reveal the practicality of the TimeSformer model for real-time deepfake detection, as it can process at a higher frames-per-second rate than the Hybrid Transformer-CNN. Further, the TimeSformer model consistently outperforms the Hybrid Transformer-CNN model on different datasets in terms of accuracy and detection rates. This model's capability to capture fine-grained temporal dynamics, especially in gaze shifts and blink patterns, lets it perform exceptionally well on datasets like FaceForensics and FakeAVCeleb, where subtle manipulations are prominent. Although it performs well for all the datasets, real differences could be brought about on specific metrics like BDA and GDE. CelebDF-V2 and DFDC have quite different pools of subjects and manipulative techniques; hence, these turn out to be challenging for both models and result in relatively low accuracy for these models in predicting blinks and gaze directions. These variations underline dataset quality, manipulation techniques, and specific characteristics of the videos to be analyzed. The robust performance of the TimeSformer model is even more remarkable in datasets with

sophisticated gaze and blink anomalies. In contrast, the ability to model such subtle spatiotemporal patterns is superior to others.

Table 12: Comparison of TimeSformer model and hybrid transformer-CNN model with published deepfake video detection methods

References	Model	FaceForensics		DFDC		CelebDF-V2		FakeAVCeleb	
		Acc. (%)	F1 Score (%)	Acc (%)	F1 Score (%)	Acc (%)	F1 Score (%)	Acc. (%)	F1 Score (%)
[75]	Xcept. (Full)	74.55	71.93	61.24	58.80	52.80	49.89	N/A	N/A
[75]	Xcept. (Face)	94.92	93.95	85.50	80.91	71.60	66.94	N/A	N/A
[35]	MesoNet	87.27	84.50	74.46	70.85	54.80	51.70	N/A	N/A
[76]	Bayar et al.	50.84	48.46	50.62	49.13	42.19	40.26	N/A	N/A
[77]	EfficientNet-b5	90.94	87.22	80.78	76.61	73.64	68.79	N/A	N/A
[78]	Inception Res.V1	79.51	76.80	59.83	57.52	45.85	43.58	N/A	N/A
[79]	Conv-LSTM	52.38	50.21	57.55	55.44	50.79	48.60	N/A	N/A
[53]	CViT	93.00	89.97	87.25	82.51	73.85	68.91	N/A	N/A
[15]	DSP-FWA	89.57	86.24	75.50	72.18	64.60	61.93	N/A	N/A
[80]	Face X-ray	79.40	76.30	65.50	62.83	54.87	50.79	N/A	N/A
[81]	3D ResNet	91.82	88.25	87.49	85.67	68.17	66.71	N/A	N/A
[82]	RECCE	92.17	90.94	88.35	85.07	72.48	71.87	N/A	N/A
[83]	RFM	90.81	86.84	87.53	84.64	70.72	69.94	N/A	N/A
[84]	DCL	92.17	91.06	89.27	84.78	75.58	72.70	N/A	N/A
[85]	FedForgery	91.38	90.05	88.08	85.51	75.74	72.87	N/A	N/A
[86]	HCiT	96.00	93.86	89.73	85.77	76.03	73.96	N/A	N/A
Our Proposed Framework	TimeSformer Model	97.5	97.4	95.8	95.7	96.3	96.3	97.1	97.2
	Hybrid Transformer-CNN Model	92.8	92.9	90.9	91.2	91.5	91.8	93.2	93.5

Overall, our study demonstrates the robustness of these frameworks on curated datasets, where the TimeSformer consistently outperformed its counterpart, thus confirming its advanced spatial-temporal modelling capability for detecting manipulations. Further, Figs. 18 and 19 present the learning curves (Training Loss and Validation Loss) for training and validation on each dataset. Figs. 20 and 21 present each dataset's learning curves (Training Score and Cross-validation Score) for training and validation.

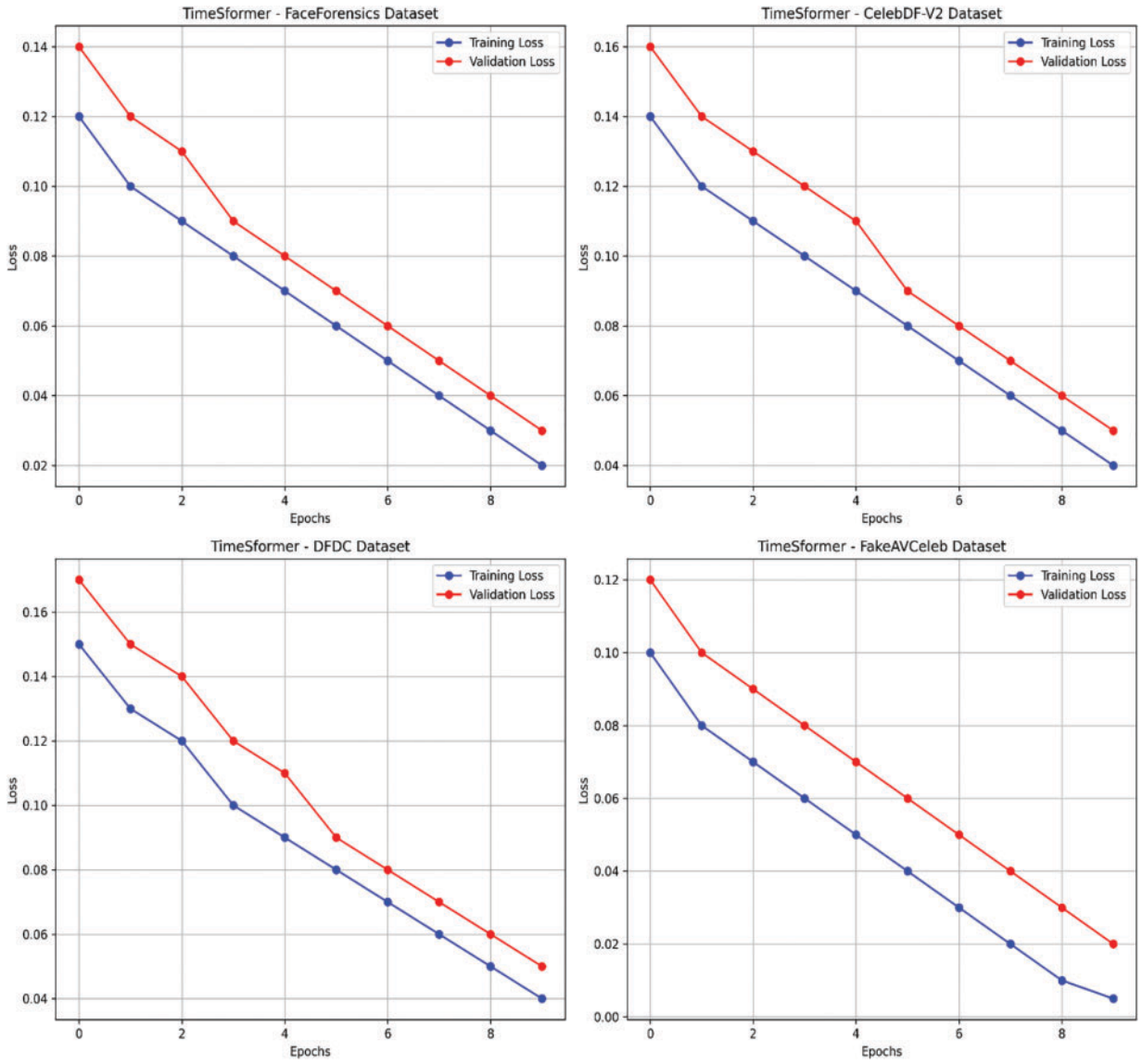

Figure 18: Learning curves (Training Loss and Validation Loss) for the TimeSformer model across different datasets

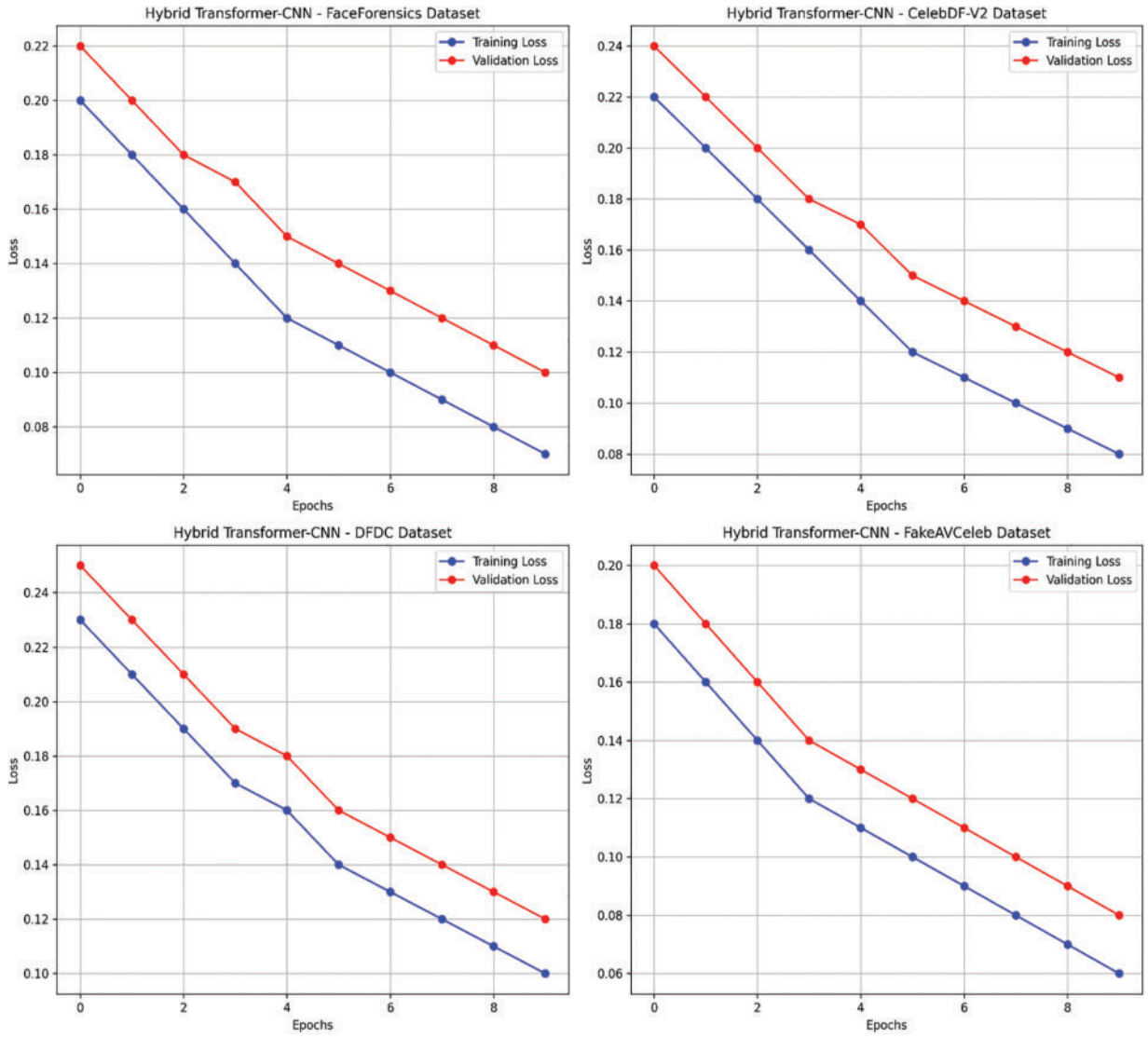

Figure 19: Learning curves (Training Loss and Validation Loss) for the hybrid transformer-CNN model across different datasets

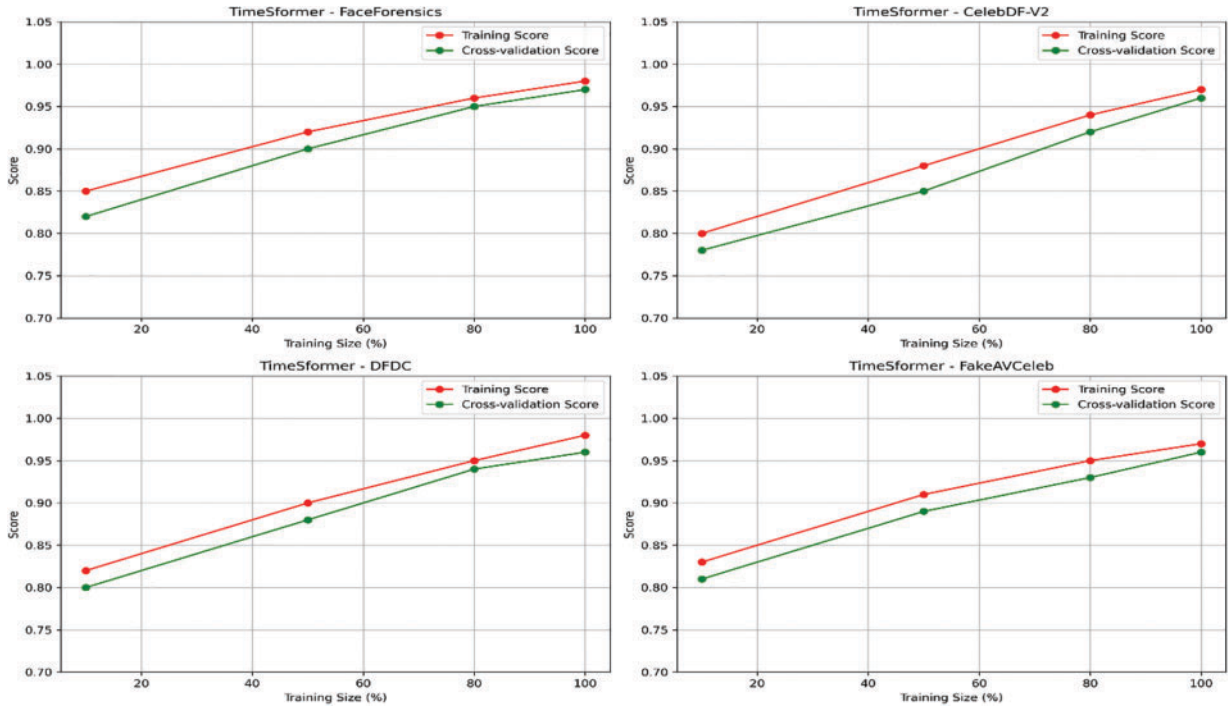

Figure 20: Learning curves (Training Score and Cross-validation Score) for the TimeSformer model across different datasets

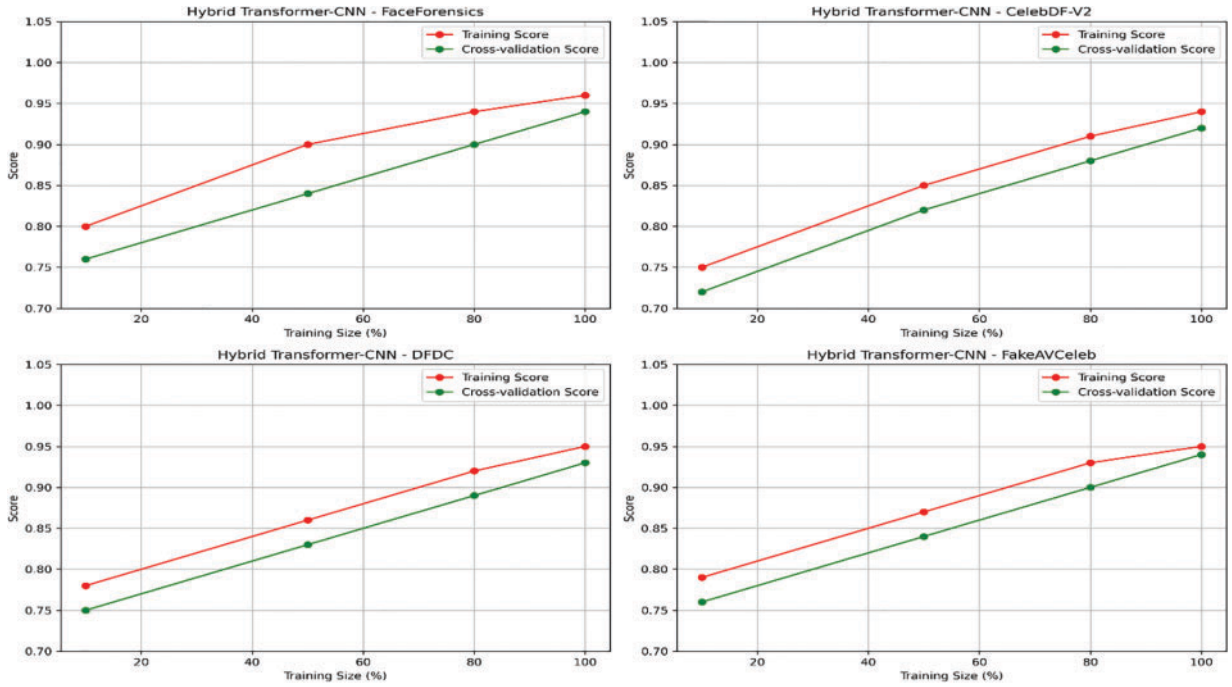

Figure 21: Learning curves (Training Score and Cross-validation Score) for the hybrid transformer-CNN model across different datasets

5 Conclusion and Future Work

Face anti-spoofing is essential for the security and integrity of face recognition systems, as it identifies and thwarts malicious attacks. Although there has been significant progress in recent years, the domain shift problem continues to pose a challenge to a model's cross-domain generalization performance. To tackle this issue, we introduced a novel framework, designed and developed based on the transformer-ss Transformer-CNN model, which incorporates MesoNet4 for lightweight spatial processing and TimeSformer for temporal modelling. Our novel framework is evaluated on four standard datasets: FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb. At the same time, the framework achieves promising results with high accuracy of 97.5%, 96.3%, 95.8%, and 97.1% on FaceForensics, CelebDF-V2, DFDC, and FakeAVCeleb datasets, respectively. It is not without limitations. One key limitation is the model's reliance on labelled datasets, which may restrict its ability to generalize to unseen types of manipulations or datasets. Additionally, the framework's performance in highly dynamic or occluded video scenarios has not been fully explored, which could impact its applicability in specific real-world environments.

Furthermore, fake face video and image detection challenges remain unresolved, including adapting models to novel and more sophisticated deepfake techniques that incorporate subtle artefacts or multimodal manipulations. To address these challenges, future work could integrate advanced domain adaptation techniques to mitigate domain shift challenges more effectively, enabling improved cross-domain generalization. Additionally, leveraging unsupervised and semi-supervised learning paradigms could help reduce dependency on labelled datasets. Exploring more lightweight transformer architectures for real-time applications and incorporating multimodal data, such as audio and thermal imaging, may further augment the robustness of face anti-spoofing systems in diverse real-world scenarios.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Muhammad Javed and Zhaohui Zhang; methodology, Muhammad Javed and Zhaohui Zhang; software, Fida Hussain Dahri; validation, Asif Ali Laghari; formal analysis, Asif Ali Laghari; investigation, writing—original draft preparation, Muhammad Javed; writing—review and editing, Martin Krajčík, Ahmad Almadhor; supervision, Zhaohui Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data is openly available in a public repository. The data that support the findings of this study are openly available in the following repositories: FaceForensics++: Available on https://paperswithcode.com/dataset/faceforensics-1 (accessed on 05 June 2025). CelebDF-V2: Available on https://paperswithcode.com/dataset/celeb-df (accessed on 05 June 2025). DFDC: Available on https://paperswithcode.com/dataset/dfdc (accessed on 05 June 2025). FakeAVCeleb: Available on https://github.com/DASH-Lab/FakeAVCeleb (accessed on 05 June 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Heidari A, Jafari Navimipour N, Dag H, Unal M. Deepfake detection using deep learning methods: a systematic and comprehensive review. Wiley Interdiscip Rev Data Min Knowl Discov. 2024;14(2):e1520. doi:10.1002/widm.1520.
2. Javed M, Zhang Z, Dahri FH, Ali Laghari A. Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach. Electronics. 2024;13(15):2947. doi:10.3390/electronics13152947.

3. Melnik A, Miasayedzenkau M, Makaravets D, Pirshtuk D, Akbulut E, Holzmann D, et al. Face generation and editing with StyleGAN: a survey. IEEE Trans Pattern Anal Mach Intell. 2024;46(5):3557–76. doi:10.1109/tpami.2024.3350004.
4. Kalpokas I, Kalpokiene J. From GANs to deepfakes: getting the characteristics right. In: Kalpokas I, Kalpokiene J, editors. Deepfakes. Cham, Switzerland: Springer International Publishing; 2022. p. 29–39. doi:10.1007/978-3-030-93802-4_4.
5. Martin-Brualla R, Radwan N, Sajjadi MSM, Barron JT, Dosovitskiy A, Duckworth D. NeRF in the wild: neural radiance fields for unconstrained photo collections. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 7206–15. doi:10.1109/cvpr46437.2021.00713.
6. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 10674–85. doi:10.1109/CVPR52688.2022.01042.
7. Malik A, Kuribayashi M, Abdullahi SM, Khan AN. DeepFake detection for human face images and videos: a survey. IEEE Access. 2022;10(1):18757–75. doi:10.1109/access.2022.3151186.
8. Guarnera L, Giudice O, Battiato S. Mastering Deepfake detection: a cutting-edge approach to distinguish GAN and diffusion-model images. ACM Trans Multimed Comput Commun Appl. 2024;20(11):1–24. doi:10.1145/3652027.
9. Khalid H, Kim M, Tariq S, Woo SS. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st Workshop on Synthetic Multimedia—Audiovisual Deepfake Generation and Detection; 2021 Oct 20–21. p. 7–15. doi:10.1145/3476099.3484315.
10. Rehaan M, Kaur N, Kingra S. Face manipulated deepfake generation and recognition approaches: a survey. Smart Sci. 2024;12(1):53–73. doi:10.1080/23080477.2023.2268380.
11. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1–11. doi:10.1109/iccv.2019.00009.
12. Xu P, Ma Z, Mei X, Shen J. Detecting facial manipulated images via one-class domain generalization. Multimed Syst. 2024;30(1):33. doi:10.1007/s00530-023-01214-7.
13. Hu C, Feng Z, Wu X, Kittler J. Dual encoder-decoder based generative adversarial networks for disentangled facial representation learning. IEEE Access. 2020;8:130159–71. doi:10.1109/access.2020.3009512.
14. Avilés-Cruz C, Celis-Escudero GJ. 3G-AN: triple-generative adversarial network under corse-medium-fine generator architecture. IEEE Access. 2023;11:105344–54. doi:10.1109/access.2023.3317897.
15. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 3204–13. doi:10.1109/cvpr42600.2020.00327.
16. Pan D, Sun L, Wang R, Zhang X, Sinnott RO. Deepfake detection through deep learning. In: Proceedings of the 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT); 2020 Dec 7–10; Leicester, UK. p. 134–43. doi:10.1109/bdcat50828.2020.00001.
17. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Appl Intell. 2023;53(4):3974–4026. doi:10.1007/s10489-022-03766-z.
18. Thies J, Zollhöfer M, Theobalt C, Stamminger M, Nießner M. *Headon*: real-time reenactment of human portrait videos. ACM Trans Graph. 2018;37(4):1–13. doi:10.1145/3197517.3201350.
19. Dahri FH, Mustafa G, Dahri U. Automatic face mask detection and recognition using deep learning. Int J Sci Eng Res. 2022;13(11):433–47. doi:10.14293/s2199-1006.1.sor-.pp7yagy.v1.
20. Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar SK. Face swapping: automatically replacing faces in photographs. In: ACM SIGGRAPH 2008 papers. Los Angeles, CA, USA: ACM; 2008. p. 1–8. doi:10.1145/1399504.1360638.
21. Saif S, Tehseen S. Deepfake videos: synthesis and detection techniques—a survey. J Intell Fuzzy Syst. 2022;42(4):2989–3009. doi:10.3233/jifs-210625.

22. Laishram L, Lee JT, Jung SK. Face de-identification using face caricature. IEEE Access. 2024;12(2):19344–54. doi:10.1109/access.2024.3356550.
23. Mukta MSH, Ahmad J, Raiaan MAK, Islam S, Azam S, Ali ME, et al. An investigation of the effectiveness of Deepfake models and tools. J Sens Actuator Netw. 2023;12(4):61. doi:10.3390/jsan12040061.
24. Fang M, Yang W, Kuijper A, Struc V, Damer N. Fairness in face presentation attack detection. Pattern Recognit. 2024;147(2):110002. doi:10.1016/j.patcog.2023.110002.
25. Leyva R, Sanchez V, Epiphaniou G, Maple C. Data-agnostic face image synthesis detection using Bayesian CNNs. Pattern Recognit Lett. 2024;183(7):64–70. doi:10.1016/j.patrec.2024.04.008.
26. Hashmi A, Shahzad SA, Ahmad W, Lin CW, Tsao Y, Wang HM. Multimodal forgery detection using ensemble learning. In: Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2022 Nov 7–10; Chiang Mai, Thailand. p. 1524–32.
27. Waseem S, Abu Bakar SARS, Ahmed BA, Omar Z, Eisa TAE, Dalam MEE. DeepFake on face and expression swap: a review. IEEE Access. 2023;11:117865–906. doi:10.1109/access.2023.3324403.
28. Khodabakhsh A, Ramachandra R, Raja K, Wasnik P, Busch C. Fake face detection methods: can they be generalized? In: Proceedings of the 2018 International Conference of the Biometrics Special Interest Group (BIOSIG); 2018 Sep 6–28; Darmstadt, Germany. p. 1–6.
29. Hoque MA, Ferdous MS, Khan M, Tarkoma S. Real, forged or deep fake? enabling the ground truth on the Internet. IEEE Access. 2021;9:160471–84. doi:10.1109/access.2021.3131517.
30. Dolhansky B. The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397v4. 2020.
31. Montserrat DM, Hao H, Yarlagadda SK, Baireddy S, Shao R, Horvath J, et al. Deepfakes detection with automatic face weighting. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14–19; Seattle, WA, USA. p. 2851–9. doi:10.1109/cvprw50498.2020.00342.
32. Fahad M, Zhang T, Iqbal Y, Ikram A, Siddiqui F, Abdullah BY, et al. Advanced deepfake detection with enhanced Resnet-18 and multilayer CNN max pooling. Vis Comput. 2025;41(5):3473–86. doi:10.1007/s00371-024-03613-x.
33. Mahum R, Irtaza A, Javed A. EDL-Det: a robust TTS synth detect using VGG19-based YAMNet ensemble learn block. IEEE Access. 2023;11:134701–16. doi:10.1109/access.2023.3332561.
34. Wei J, Lu G, Liu H, Yan J. Facial image inpainting with deep generative model and patch search using region weight. IEEE Access. 2019;7:67456–68. doi:10.1109/access.2019.2919169.
35. Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: a compact facial video forgery detection network. In: Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China. p. 1–7. doi:10.1109/WIFS.2018.8630761.
36. Yang S, Huang H, Huang YS, Jin X. Facial action units detection using temporal context and feature reassignment. Comput Animat Virtual. 2024;35(3):e2246. doi:10.1002/cav.2246.
37. Ali Channa I, Li D, Dahri FH, Mustafa Abro GE, Zahid F. A novel deep learning model for classifying power quality problems in PV-integrated microgrids using CNN-LSTM. In: Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR); 2024 May 14–15; Muscat, Oman. p. 1–5. doi:10.1109/ICIESTR60916.2024.10798309.
38. Jung T, Kim S, Kim K. Deep: deep detect using hum eye blinking pattern. IEEE Access. 2020;8:83144–54. doi:10.1109/access.2020.2988660.
39. Ilyas H, Javed A, Malik KM. AVFakeNet: a unified end-to-end Dense Swin Transformer deep learning model for audio—visual deepfakes detection. Appl Soft Comput. 2023;136(11):110124. doi:10.1016/j.asoc.2023.110124.
40. Zhang R, Wang H, Du M, Liu H, Zhou Y, Zeng Q. UMMAFormer: a universal multimodal-adaptive transformer framework for temporal forgery localization. In: Proceedings of the 31st ACM International Conference on Multimedia; 2023 Oct 27–31; Ottawa, ON, Canada. p. 8749–59. doi:10.1145/3581783.3613767.
41. Sheng B, Li P, Ali R, Philip Chen CL. Improving video temporal consistency via broad learning system. IEEE Trans Cybern. 2022;52(7):6662–75. doi:10.1109/tcyb.2021.3079311.
42. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? Proc Mach Learn Res. 2021;139:813–24.

43. Khaled N, Saad S, Aref M. TimeSformer-MIL: a hybrid approach for anomalous activity recognition in real-world surveillance videos. In: Proceedings of the 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES); 2024 Oct 19–21. Giza, Egypt. p. 139–42. doi:10.1109/NILES63360.2024.10753147.
44. Khalid H, Tariq S, Kim M, Woo SS. FakeAVCeleb: a novel audio-video multimodal deepfake dataset. arXiv:2108.05080v4. 2022.
45. Javed M, Zhang Z, Dahri FH, Kumar T. Enhancing multimodal deepfake detection with local-global feature integration and diffusion models. Signal Image Video Process. 2025;19(5):400. doi:10.1007/s11760-025-03970-7.
46. Yu Z, Cai R, Li Z, Yang W, Shi J, Kot AC. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE Trans Dependable Secur Comput. 2024;21(5):4327–42. doi:10.1109/tdsc.2024.3352049.
47. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 2625–34. doi:10.1109/CVPR.2015.7298878.
48. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
49. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 2261–9. doi:10.1109/CVPR.2017.243.
50. Ciftci UA, Demir I, Yin L. FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. 2024:1. doi:10.1109/tpami.2020.3009287.
51. Tarasiou M, Zafeiriou S. Extracting deep local features to detect manipulated images of human faces. In: Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP); 2020 Oct 25–28; Abu Dhabi, United Arab Emirates. p. 1821–5. doi:10.1109/icip40778.2020.9190714.
52. Coccomini DA, Messina N, Gennaro C, Falchi F. Combining EfficientNet and vision transformers for video deepfake detection. In: International Conference on Image Analysis and Processing. Cham, Switzerland: Springer International Publishing; 2022. p. 219–29. doi:10.1007/978-3-031-06433-3_19.
53. Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer. arXiv:2102.11126v3. 2021.
54. Pintelas E, Pintelas P. A 3D-CAE-CNN model for deep representation learning of 3D images. Eng Appl Artif Intell. 2022;113(2):104978. doi:10.1016/j.engappai.2022.104978.
55. Zhao C, Wang C, Hu G, Chen H, Liu C, Tang J. ISTVT: interpretable spatial-temporal video transformer for deepfake detection. IEEE Trans Inform Forensic Secur. 2023;18(1):1335–48. doi:10.1109/tifs.2023.3239223.
56. Yin Q, Lu W, Li B, Huang J. Dynamic difference learning with spatio-temporal correlation for deepfake video detection. IEEE Trans Inform Forensic Secur. 2023;18:4046–58. doi:10.1109/tifs.2023.3290752.
57. Ismail A, Elpeltagy M, Zaki MS, Eldahshan K. An integrated spatiotemporal-based methodology for deepfake detection. Neural Comput Appl. 2022;34(24):21777–91. doi:10.1007/s00521-022-07633-3.
58. de Lima O, Franklin S, Basu S, Karwoski B, George A. Deepfake detection using spatiotemporal convolutional networks. arXiv:2006.14749v1. 2020.
59. Guo JM, Yang JS, Seshathiri S, Wu HW. A light-weight CNN for object detection with sparse model and knowledge distillation. Electronics. 2022;11(4):575. doi:10.3390/electronics11040575.
60. Pokroy AA, Egorov AD. EfficientNets for DeepFake detection: comparison of pretrained models. In: Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus); 2021 Jan 26–29; Petersburg, Moscow, Russia. p. 598–600. doi:10.1109/elconrus51938.2021.9396092.
61. Khan SA, Dai H. Video transformer for deepfake detection with incremental learning. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Online. doi:10.1145/3474085.3475332.
62. Guo J, Zhu X, Yang Y, Yang F, Lei Z, Li SZ. Towards fast, accurate and stable 3D dense face alignment. In: Proceedings of the Computer Vision-ECCV 2020; Glasgow, UK. p. 152–68. doi:10.1007/978-3-030-58529-7_10.
63. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.

64. Dufour N, Gully A. Contributing data to deepfake detection research. Google AI Blog. 2019;1(2):3.
65. Haliassos A, Vougioukas K, Petridis S, Pantic M. Lips don't lie: a generalisable and robust approach to face forgery detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 5037–47. doi:10.1109/cvpr46437.2021.00500.
66. Li Y, Chang MC, Lyu S. In ictu oculi: exposing AI created fake videos by detecting eye blinking. In: Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China. p. 1–7. doi:10.1109/WIFS.2018.8630787.
67. Amerini I, Galteri L, Caldelli R, Del Bimbo A. Deepfake video detection through optical flow based CNN. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27–28; Seoul, Republic of Korea. p. 1205–7. doi:10.1109/iccvw.2019.00152.
68. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D. Emotions don't lie: an audio-visual deepfake detection method using affective cues. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. p. 2823–32. doi:10.1145/3394171.3413570.
69. Haliassos A, Mira R, Petridis S, Pantic M. Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 18–24; New Orleans, LA, USA. p. 14930–42. doi:10.1109/CVPR52688.2022.01453.
70. Zheng Y, Bao J, Chen D, Zeng M, Wen F. Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 15024–34. doi:10.1109/ICCV48922.2021.01477.
71. Wang J, Wu Z, Ouyang W, Han X, Chen J, Jiang YG, et al. M2TR: multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 International Conference on Multimedia Retrieval; 2020 Jun 27–30; Newark, NJ, USA. p. 615–23. doi:10.1145/3512527.3531415.
72. Li Y, Chang MC, Lyu S. In ictu oculi: exposing AI generated fake face videos by detecting eye blinking. arXiv:1806.02877. 2018.
73. Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs. Int J Comput Vis. 2020;128(5):1398–413. doi:10.1007/s11263-019-01251-8.
74. Pham HX, Wang Y, Pavlovic V. Generative adversarial talking head: bringing portraits to life with a weakly supervised neural network. arXiv:1803.07716. 2018.
75. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1800–7. doi:10.1109/CVPR.2017.195.
76. Bayar B, Stamm MC. A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security; 2015 Jun 20–22; Vigo Galicia, Spain. p. 5–10. doi:10.1145/2909827.2930786.
77. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning; 2019 Jun 10–15; Long Beach, CA, USA. p. 6105–14.
78. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA. doi:10.1609/aaai.v31i1.11231.
79. Güera D, Delp EJ. Deepfake video detection using recurrent neural networks. In: Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); 2018 Nov 27–30; Auckland, New Zealand. p. 1–6. doi:10.1109/AVSS.2018.8639163.
80. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, et al. Face X-ray for more general face forgery detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 5000–9. doi:10.1109/cvpr42600.2020.00505.
81. Roy R, Joshi I, Das A, Dantcheva A. 3D CNN architectures and attention mechanisms for deepfake detection. In: Rathgeb C, Tolosana R, Vera-Rodriguez R, Busch C, editors. Handbook of digital face manipulation and detection. Cham, Switzerland: Springer International Publishing; 2022. p. 213–34. doi:10.1007/978-3-030-87664-7_10.

82. Cao J, Ma C, Yao T, Chen S, Ding S, Yang X. End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 4103–12. doi:10.1109/CVPR52688.2022.00408.
83. Wang C, Deng W. Representative forgery mining for fake face detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 14918–27. doi:10.1109/cvpr46437.2021.01468.
84. Sun K, Yao T, Chen S, Ding S, Li J, Ji R. Dual contrastive learning for general face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2022 Feb 22–Mar 1. p. 2316–24. doi:10.1609/aaai.v36i2.20130.
85. Liu D, Dang Z, Peng C, Zheng Y, Li S, Wang N, et al. FedForgery: generalized face forgery detection with residual federated learning. IEEE Trans Inform Forensic Secur. 2023;18(1):4272–84. doi:10.1109/tifs.2023.3293951.
86. Kaddar B, Fezza SA, Akhtar Z, Hamidouche W, Hadid A, Serra-Sagristá J. Deepfake detection using spatiotemporal transformer. ACM Trans Multimed Comput Commun Appl. 2024;20(11):1–21. doi:10.1145/3643030.