Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.084565

ARTICLE

HCBLNet: An Attention Enhanced Hybrid CNN-BiLSTM Network Model for Deepfake Video Detection with Explainable Artificial Intelligence Interpretation

Ashifur Rahman[1,*], Md. Rajaul Karim[2], Amran Hossain[3,*], Abdulmajeed M. Alenezi[4,*], Md. Shabiul Islam[5,6,*] and Mohammad Tariqul Islam[7,8]

[1]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[2]School of Computing, Georgia Southern University, Statesboro, GA, USA
[3]Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Gazipur, Bangladesh
[4]Department of Electrical Engineering, Faculty of Engineering, Islamic University of Madinah, Madinah, Saudi Arabia
[5]Centre for Advanced Devices and Systems, Centre of Excellence for Robotics and Sensing Technologies, Multimedia University, Cyberjaya, Malaysia
[6]Faculty of Artificial Intelligence and Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia
[7]Department of Electrical Engineering, College of Engineering, Taif University, Taif, Saudi Arabia
[8]Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia

*Corresponding Authors: Ashifur Rahman. Email: ashifurrahman.bubt@gmail.com; Amran Hossain. Email: amran.cse@duet.ac.bd; Abdulmajeed M. Alenezi. Email: a.alenezi@iu.edu.sa; Md. Shabiul Islam. Email: shabiul.islam@mmu.edu.my

Received: 25 April 2026; Accepted: 07 September 2026; Published: 21 October 2026

ABSTRACT: Deepfake videos pose an increasing societal threat by enabling the spread of deceptive content, shaping public opinion, and undermining trust in digital media. Detecting such manipulated content is essential to preserving the credibility of online information. Although recent deepfake detection methods have shown promising progress, many struggle to identify highly sophisticated forgeries and generalize effectively across diverse datasets. To address these challenges, this paper proposes a deepfake detection framework, termed HCBLNet, which integrates multi-head attention with a hybrid CNN–BiLSTM architecture. A pre-trained ResNet50 is employed for robust spatial feature extraction, capturing fine-grained visual artifacts indicative of deepfake manipulations. The incorporation of multi-head attention enhances the modeling of spatial dependencies among regions within each frame, while the BiLSTM component subsequently learns sequential patterns across these enhanced spatial features, improving frame-level detection robustness. To enhance interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) are applied to provide visual explanations by highlighting manipulated regions in video frames. The proposed model is evaluated on three widely used benchmark datasets: Deepfake Detection Challenge (DFDC), FaceForensics++, and Celeb-DF. Experimental results demonstrate strong and consistent performance, achieving accuracies of 97.91% on DFDC, 95.40% on FaceForensics++, and 95.97% on Celeb-DF. Furthermore, we compare our model with seven popular transfer learning models, including VGG16, VGG19, ResNet50, ResNet152, DenseNet121, DenseNet201, and EfficientNet, where HCBLNet consistently achieves higher performance across all evaluation metrics under the experimental conditions considered in this study. These results indicate that HCBLNet is a promising and effective hybrid approach for deepfake detection on the evaluated benchmarks. Overall, the proposed approach offers an accurate and interpretable solution for deepfake detection, with potential for applications in media forensics, content moderation, journalism, and cybersecurity.

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

KEYWORDS: Deepfake video detection; convolutional neural network (CNN); bidirectional long short-term memory (Bi-LSTM); multi-head attention; deepfake detection challenge dataset (DFDC); FaceForensics++

1 Introduction

In recent years, we have witnessed a remarkable technological revolution in digital media creation and manipulation. Rapid developments in machine learning (ML), computer vision, and graphics processing technologies enabled broader content generation, allowing individuals from diverse backgrounds to access it [1,2]. However, this accessibility has also led to the proliferation of manipulated media, e.g., deepfake videos, images, or audios [3]. Deepfake technology leverages artificial intelligence, specifically deep learning, to create manipulated videos. These videos often depict actions performed by individuals, but with faces swapped to those of different people through advanced AI techniques [1,4,5]. Although deepfake technology has some ethical uses, such as in the entertainment sector and in education, it's regrettable that it's also misused for harmful purposes such as blackmailing individuals, committing financial fraud, etc. [6]. Public figures are particularly vulnerable to deepfakes because of the vast availability of their visual content across online platforms [7]. Two prominent deepfake cases are frequently seen on social media. The first involves deepfake technology in pornography, and the second concerns a political smear campaign [8]. In the first scenario, deepfake technology is used to replace the actors' faces in pornographic adult content with those of celebrities or well-known individuals, aiming to damage their reputations. In the second scenario, videos are altered to show certain individuals making controversial statements, with their faces swapped with those of political figures involved in elections, intending to harm their public image and reduce their electability [9].

Creating deepfake content involves the use of advanced neural networks such as Generative Adversarial Networks (GANs) and Deep Autoencoders [10]. The Autoencoders contain two parts: an encoder, which turns an image into data, and a decoder, which provides a reconstructed image from that data. To swap faces, two Autoencoders are used with a shared encoder but different decoders [10]. GANs work by having a generator create fake content and a discriminator detect if the content is real or fake [11]. Although these methods can produce very realistic deepfakes, they have detectable flaws, such as differences in image resolution, mismatched facial styles with body parts, and unnatural facial movements [12,13]. The deepfake detection method aims to distinguish between deepfake and original content in images or videos through a binary classification process [8]. It involves extracting features from the content to differentiate between the two classes [14]. There are four main types of deepfake detection methods based on feature extraction approaches [8]: visual feature-based, which uses observable features like blinking and facial organ shapes; local feature-based, which analyzes pixel-level features for higher reliability than visual methods; deep feature-based, which extracts complex features using multiple layers at the pixel level; and temporal feature-based, which captures transient features from consecutive frames, applicable exclusively to video content.

Deep learning models act as a significant component in recognizing fake videos because they can process and interpret intricate structures within visual data [1]. These architectures, such as Long Short-Term Memory Networks (LSTMs), Bidirectional LSTMs (BiLSTMs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs), can identify subtle inconsistencies and artifacts that are often present in manipulated media [15,16]. Additionally, transfer learning models, such as ResNet, DenseNet, EfficientNet, and VGG-Net also used to detect deepfake videos in different research [8,17]. In a similar vein, a hybrid CNN-LSTM (EfficientNet-LSTM) architecture has been proposed for deepfake (object forgery) detection in videos, where EfficientNet extracts spatial features from individual frames while LSTM models the temporal dependencies across frames, allowing the joint detection of spatial and temporal manipulation

artifacts [18]. On the other hand, the integration of attention mechanisms with CNN and LSTM has recently been applied to various image classification tasks. The attention mechanism enhances model performance by focusing on important regions within frames, aiding in anomaly detection. To further improve interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated [19] in different architectures. Grad-CAM produces heatmaps that visually highlight important regions in video frames that contribute most to the deepfake classification decision, providing insight into how the model arrives at its conclusions and assisting in uncovering potential bias or incorrect predictions. By utilizing extensive video datasets containing both fabricated and real footage, these architectures learn to distinguish between authentic footage and deepfakes based on anomalies in facial expressions, skin textures, lighting inconsistencies, and temporal discrepancies [20]. Various benchmark datasets, such as Deepfake Detection Challenge (DFDC) [20], FaceForensics++ [21], and Celeb-DF [22], are widely utilized for different purposes of deepfake detection and analysis.

In general, while numerous strategies have been proposed to detect deepfakes, their effectiveness often diminishes when applied to unfamiliar or novel video content. The growing sophistication of deepfake generation, powered by advanced deep learning algorithms, further complicates detection efforts, considering factors such as variations in lighting conditions, image compression, scaling, and subject positioning that introduce additional challenges. As a result, the increasing misuse of deepfakes in deceptive practices has motivated the research community to explore more robust and generalizable identification approaches [23]. Addressing these issues and the limitations of earlier efforts, we introduced HCBLNet, a deepfake detection method combining multi-head attention with a hybrid CNN-BiLSTM architecture. Using a pre-trained ResNet50 for feature extraction, the model captures intricate visual details crucial for identifying deepfakes in DFDC, FaceForensics++, and Celeb-DF datasets. The multi-head attention mechanism allows the model to attend to different spatial regions within each frame at the same time, helping it grasp diverse contextual relationships among these regions, while the BiLSTM subsequently models dependencies among the resulting spatial features to produce a frame-level real/fake classification.

This paper presents the following key contributions:

- We introduced a deep learning based hybrid architecture called HCBLNet that combines Bi-LSTM and CNN with multi-head attention for efficient frame-level deepfake detection with improved accuracy.
- The proposed HCBLNet model enhanced feature extraction capability and reliability, showing strong robustness compared with the transfer learning baselines considered in this study.
- The overall performance of the proposed HCBLNet model is evaluated by utilizing the diverse benchmark datasets across DFDC, FaceForensics++, and Celeb-DF to assess its performance and generalization behavior on visual forgery detection.
- We emphasize the significance of BiLSTM architectures in effectively modeling sequential dependencies among spatial features extracted from individual frames for deepfake detection.
- We highlight how the attention mechanism applied to spatial regions within a frame improves the accuracy of deepfake detection models.
- We analyze the prediction outcomes of the model by applying the Grad-CAM and Local Interpretable Model-Agnostic Explanations (LIME) explainable AI (XAI) technique to produce visual insights enabling greater transparency and interpretability in deepfake detection.
- Finally, the proposed HCBLNet model is compared with transfer learning based deep learning models and prior deepfake approaches to assess its effectiveness and detection capabilities under the experimental conditions considered in this study.

The structure of the paper is outlined as follows: Section 2 offers an in-depth overview of prior research conducted in this domain. Section 3 covers materials and methods, while Section 4 discusses all the results and comparisons with existing literature. Section 5 presents an in-depth analysis of the proposed model, focusing on its advantages and potential drawbacks, while Section 6 concludes the paper and outlines future research directions.

2 Related Work

Deepfake detection remains an ongoing challenge, and continuous advancement of detection techniques is necessary to stay ahead of the curve and protect against misinformation and deception [15]. Beyond deepfake videos, artificial intelligence techniques have also been extensively studied for combating fake news on social media, encompassing text-based, image-based, deepfake video-based, and multimodal misinformation detection approaches [24]. Deepfake technology involves manipulating images and videos to create convincing yet fabricated content. Prior to the emergence of deepfake technology, conventional approaches for altering visual media involved practices such as splicing of the images, where objects are overwritten within an image, either from the same image or different ones, resulting in additions, duplications, shifts, or removals of objects [25]. Similarly, video splicing involves altering videos by inserting or removing frames to achieve specific effects. These practices collectively fall under the umbrella of copy-move forgery. Detecting such manipulations typically involves using feature extraction algorithms to identify duplicated parts in images and videos through matching processes [26–30]. Another approach to detecting image and video splicing is through statistical analysis, which can reveal anomalies in the content of images and videos [31–33].

Deep learning models are essential tools in detecting deepfake videos as they can evaluate and identify intricate structures and irregularities that typically go unnoticed by human observation. The study [34] focused on building a deepfake detection system by integrating LSTM with a VGG-16 network. It conducted a comparative analysis of ResNet-101, VGG-16, and VGG-19 architectures, each paired with LSTM, using the Celeb-DF video dataset. Among the evaluated models, the VGG-16 configuration trained over 15 epochs with a mini-batch size of 32 delivered the best performance, reaching an accuracy of approximately 96%, a recall close to 93%, specificity near 99.2%, and precision around 99%.

The PUDD [35], a deepfake detection framework leveraging prototype-based learning principles that achieves 95.1% accuracy on Celeb-DF. PUDD leverages prototype similarity for video classification, demonstrating efficient retraining in just 2.7 s. In addition, paper [36] addresses concerns about synthetic image impact by proposing a deep-learning method for detecting fraudulent videos. Their ResNet-Swish-BiLSTM hybrid approach achieves 96.23% accuracy on FF++ and 78.33% on combined FF++ and DFDC datasets in real-time deepfake detection.

Authors of the paper [37] introduce an ensemble classification approach based on deep learning and bagging techniques, leveraging the CoAtNet model for identifying tampered facial content in video data. This method combines depthwise convolutions with self-attention mechanisms to improve the model's ability to generalize while maintaining computational effectiveness. Trained on randomly sampled subsets of data and augmented with CutMix, their method achieves accuracy on FaceForensics++: NeuralTextures (NT), DeepFakes (DF), FaceSwap (FS), Face2Face (F2F); 97.7%, 91.69%, 95.74%, and 79.32%, respectively, along with 97.81% on Celeb-DF. In the paper [38], the authors present a deepfake identification method built upon the Vision Transformer (ViT) framework, focusing on distinguishing between deepfake and authentic videos. Their approach utilizes spatiotemporal features extracted from landmark area images using ViT, enhanced by convolution blocks that are separable depthwise and an attention module of the convolution

block. Training the Celeb-DF V.2 dataset with augmentation, their system achieved 87.18% accuracy and a 92.52% F1 score after 100 epochs with the Adam optimizer.

Identification of lip-forgery by exploiting inconsistencies in the movements of lip with the auditory input poses another constraint in deep-fake recognition. The authors of [39] propose a novel approach focused on detecting lip-forgery by exploiting inconsistencies between lip movements and audio signals and integrating biological cues between lips and head regions. Using a high-quality Lip-Sync dataset created with advanced lip generation techniques, their method achieves over 95.3% accuracy in detecting lip-syncing videos, outperforming existing methods. It demonstrates robustness across diverse inputs and achieves accuracy approximative 90% in realistic contexts, underscoring its practical deployment capabilities.

In the paper [40], the author proposes a hybrid architecture that leverages a Convolutional Vision Transformer to identify deepfake content, where Convolutional Neural Networks (CNNs) are utilized for extracting features and Vision Transformers (ViTs) are employed for classification through attention-based mechanisms. Trained on the DFDC dataset, their model reports an accuracy of 91.5%, an AUC score of 0.91, and a loss value of 0.32, highlighting its strong performance achieved through the incorporation of CNNs within a Vision Transformer-based framework. Additionally, authors of [41] introduce Facial Action Dependency Estimation (FADE), framing deepfake detection as a graph classification task with the Multi-Dependency Graph Module (MDGM) that extracts the dependencies associated with facial action, significantly enhancing detection performance accuracy to 92.89%. In another study [42] presented a hybrid deep learning framework, InceptionResNet-BiLSTM, to detect deepfake videos effectively despite variations in generation techniques, illumination, and ethnicities. The model combines a customized InceptionResNetV2 for frame-level feature extraction and a BiLSTM for temporal classification. Evaluated on the diverse FaceForensics++ and FakeAVCeleb datasets, their method achieves over 90% accuracy across datasets such as FaceSwap, Face2Face, and DeepFakes, indicating strong effectiveness and adaptability to various manipulation techniques.

In the study [43], the authors introduce NoiseDF, a noise-based detection model leveraging forensic traces in Deepfake videos. They extend the capabilities of RIDNet to learn noise patterns from facial and background regions, and propose a Multi-Head Relative-Interaction mechanism to support critical evaluations in deepfake detection. Outperforming state-of-the-art models, NoiseDF visualizes forensic noise traces, robustly demonstrating its effectiveness in Deepfake detection with 84.36% on FF++, 59.87% on DFDC, and 70.10% on the Celeb-DF dataset. Additionally, authors of [44] introduce ResViT, a model combining ResNet for feature extraction and a vision transformer for classification, to classify fake visual content efficiently. Through the analysis of visual frames and a strong focus on pre-processing techniques, ResViT enhances performance. Extensive experiments on five datasets show ResViT achieving prediction accuracies of 80.48% on Celeb-DF, 87.23% on Celeb-DFv2, 75.62% on FaceForensics++, 78.45% on FF-Deepfake Detection, and 84.55% on DFDC.

A major challenge in DeepFake detection is the poor generalization of state-of-the-art algorithms across different facial manipulations. The authors of [45] introduce ID-Reveal, a DeepFake detection approach that learns temporal facial features using metric learning and adversarial training with only real videos. Utilizing high-level semantic features, ID-Reveal is robust to post-processing and shows significant improvement in generalization. Additionally, authors of [46] present a deep learning framework based on a convolution-enhanced Transformer architecture for detecting DeepFakes, which combines both localized and holistic visual representations through convolutional pooling and a refined attention mechanism. By utilizing keyframes for training, their approach surpasses existing models across multiple evaluation datasets, showcasing strong results in both intra- and inter-dataset scenarios.

Overall, previous studies have experimented with various models on different DeepFake datasets such as FaceForensics++, DFDC, DFD, DF-1.0, LRS2, Celeb-DF, etc. Most methods were tested on one or two datasets, showing minimal comparison with existing works and achieving relatively low accuracy. In addition, a major limitation of many existing approaches is the lack of interpretability. While models achieve high accuracy, they often fail to provide visual explanations for their predictions, making it difficult to understand which regions of the face contributed most to the decision. This black-box nature hinders trust and real-world applicability. Grad-CAM, which is a visual explanation method that emphasizes the most influential regions within input frames, is rarely integrated with current models, leaving a gap in explainability. To address these limitations, the proposed method employs a unique architecture, incorporates Grad-CAM for enhanced interpretability, demonstrates superior results across different datasets, and is compared with the majority of recent works. A comparison table, including references, datasets, and accuracy of existing studies alongside our proposed model, is provided in the Section 4.8.

3 Materials and Methods

This section presents a detailed methodology developed for identifying deepfake videos, emphasizing the processes of data preprocessing, network design, and performance assessment. A graphical abstract of the study pipeline is presented in Fig. 1. The methodology is designed to leverage advanced techniques in computer vision and machine learning to effectively identify manipulated video content. Details of each subsection are provided subsequently.

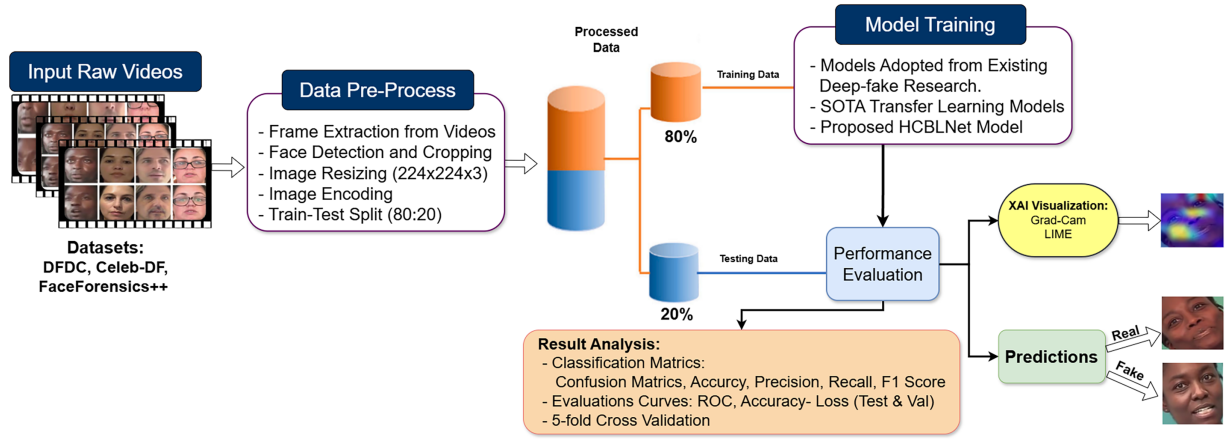


Figure 1: Comprehensive pipeline illustrating the entire study.

3.1 Dataset Description

We used three publicly available benchmark datasets for the proposed HCBLNet experiments: DFDC [20], FaceForensics++ [21], and Celeb-DF [22]. Each dataset consists of two classes, real and fake videos, providing a comprehensive basis for evaluating deepfake detection performance. From the original DFDC dataset, which consists of multiple officially released subsets provided by the dataset creators, we utilized the first available subset in the DFDC release. This subset contains 400 videos, comprising 200 real and 200 fake samples, which naturally maintains class balance at the video level. The use of this predefined subset avoids any ad-hoc sampling procedure and ensures consistency with the dataset's original partitioning structure. This selection strategy also preserves the inherent distribution of the dataset without introducing additional selection bias. The FaceForensics++ and Celeb-DF datasets were used following their standard configurations, which contain high-quality and realistically manipulated videos.

It is also important to note that the selected benchmark datasets (DFDC, FaceForensics++, and Celeb-DF) are widely recognized in the deepfake detection literature and are constructed to reflect real-world and in-the-wild conditions [47,48]. Therefore, they inherently include various challenging distortions such as video compression artifacts, variations in illumination, differences in resolution, motion blur, background clutter, and encoding noise. As a result, these datasets already provide a realistic evaluation setting that partially covers robustness scenarios typically observed in practical deepfake content.

To enable frame-based learning, videos from all datasets were uniformly decomposed into frames at a fixed sampling rate of one frame every 10 frames. This strategy balances temporal coverage and computational efficiency while ensuring consistent frame extraction across datasets. As a result, each video contributes a variable number of frames depending on its duration. Fig. 2 illustrates representative sample frames extracted from the datasets used in this study.



Figure 2: Sample frames of real and deepfake videos collected from the datasets.

The observed difference in the number of real and fake frames arises from variations in video lengths and durations rather than intentional class imbalance. To mitigate any bias introduced at the frame level, dataset splitting and evaluation were performed at the video level rather than the frame level. All datasets were divided into training and testing sets using an 80:20 split at the video level, ensuring that frames extracted from the same video do not appear in both sets. This prevents data leakage and supports fair performance evaluation. Table 1 summarizes the number of videos and extracted frames for each dataset, reflecting the resulting distribution used in our experiments.

Table 1: Data distribution table across different datasets used in the proposed method.

Dataset	Total Videos	Train Videos	Test Videos	Real Frames	Fake Frames	Total Frames
DFDC	400	320	80	760	915	1675
FaceForensics++	1000	800	200	5605	6028	11,633
Celeb-DF	6529	5223	1306	7813	11,278	19,091

3.2 Data Preprocessing

The three datasets employed for this study comprise videos containing human faces, sourced to ensure diversity in facial features, lighting, and backgrounds. Each video is split into individual frames using

OpenCV [49] for detailed frame-by-frame analysis. From each frame, faces are detected and cropped using the *face_recognition* [50] library of Python. The region of interest is extracted for each detected face, retaining only the relevant portions of the frames. These cropped faces are then saved as individual image files in two separate directories, 'real faces' and 'fake faces', and are utilized to test and train the proposed model. These cropped faces are resized to a uniform dimension of $224 \times 224 \times 3$ and converted into arrays for compatibility with the model. One-hot encoding is applied to label each image according to its class. The dataset is split in an 80:20 ratio, with 80% used for training the model and 20% reserved for testing, allowing for effective learning and reliable performance evaluation. To further assess the model's robustness and generalizability to new data, 5-fold cross-validation is also employed.

3.3 Proposed Architecture

The proposed HCBLNet architecture is depicted in Fig. 3. The overall pipeline includes four components: "Feature Extraction with ResNet50", "Spatial Feature Enhancement with Multi-Head Attention", "Sequence Modeling with Bidirectional LSTM", and "Interpretability with Grad-CAM and LIME Visualization". The Grad-CAM module is incorporated to produce class-specific visual interpretations by emphasizing key areas within video frames that impact the model's predictions, thereby improving interpretability and trust in the decision-making process. Complementarily, the LIME module provides local, model-agnostic explanations by identifying superpixel-level contributions to individual predictions. Detailed descriptions of each of these components are provided in the subsequent sections. We note that HCBLNet operates at the frame level: each video frame is independently extracted, processed, and classified as real or fake, and the model does not construct or aggregate multi-frame clips.

3.3.1 Feature Extraction with ResNet50

This utilizes ResNet50, a deep residual network, to derive advanced spatial representations from individual video frames. The pre-trained ResNet50 architecture on the ImageNet dataset is leveraged with the exclusion of the final classification layer. This pre-trained model processes input images of size $224 \times 224 \times 3$. The backbone further analyzes these input images through multiple convolutional and identity blocks, thereby extracting hierarchical features. The resultant output is a feature map of shape $F \in \mathbb{R}^{7\times7\times2048}$. This approach allows the utilize rich pre-trained features while simultaneously reducing computational overhead.

Here, the input image $I \in \mathbb{R}^{224\times224\times3}$ is processed to produce a feature map $F \in \mathbb{R}^{7\times7\times2048}$.

$$F = \text{ResNet50}(I) \tag{1}$$

3.3.2 Spatial Feature Enhancement with Multi-Head Attention

To enhance the discriminative spatial representation extracted from each individual frame, we employed a multi-head self-attention strategy that enables the model to attend to different spatial regions of the same frame in parallel. Since ResNet50 produces a 7×7 grid of spatial locations for each frame, this attention mechanism models the relationships and dependencies among these 49 spatial positions within a single frame—it does not operate across different frames or capture temporal information. Temporal reasoning, where applicable, is not part of this stage; HCBLNet classifies each frame independently, and this module's role is purely to refine the spatial feature representation of that frame before it is passed to the sequence modeling stage described in Section 3.3.3. In this approach, the feature maps produced by ResNet50, initially with dimensions of $7 \times 7 \times 2048$, are reshaped into a sequential format of 49×2048, where each of the 49 elements corresponds to one spatial location (patch) of the 7×7 grid within the frame,

and the 2048 dimension represents the feature channels at that location. Within the multi-head attention mechanism, the model employs 8 attention heads, each executing the following operations:

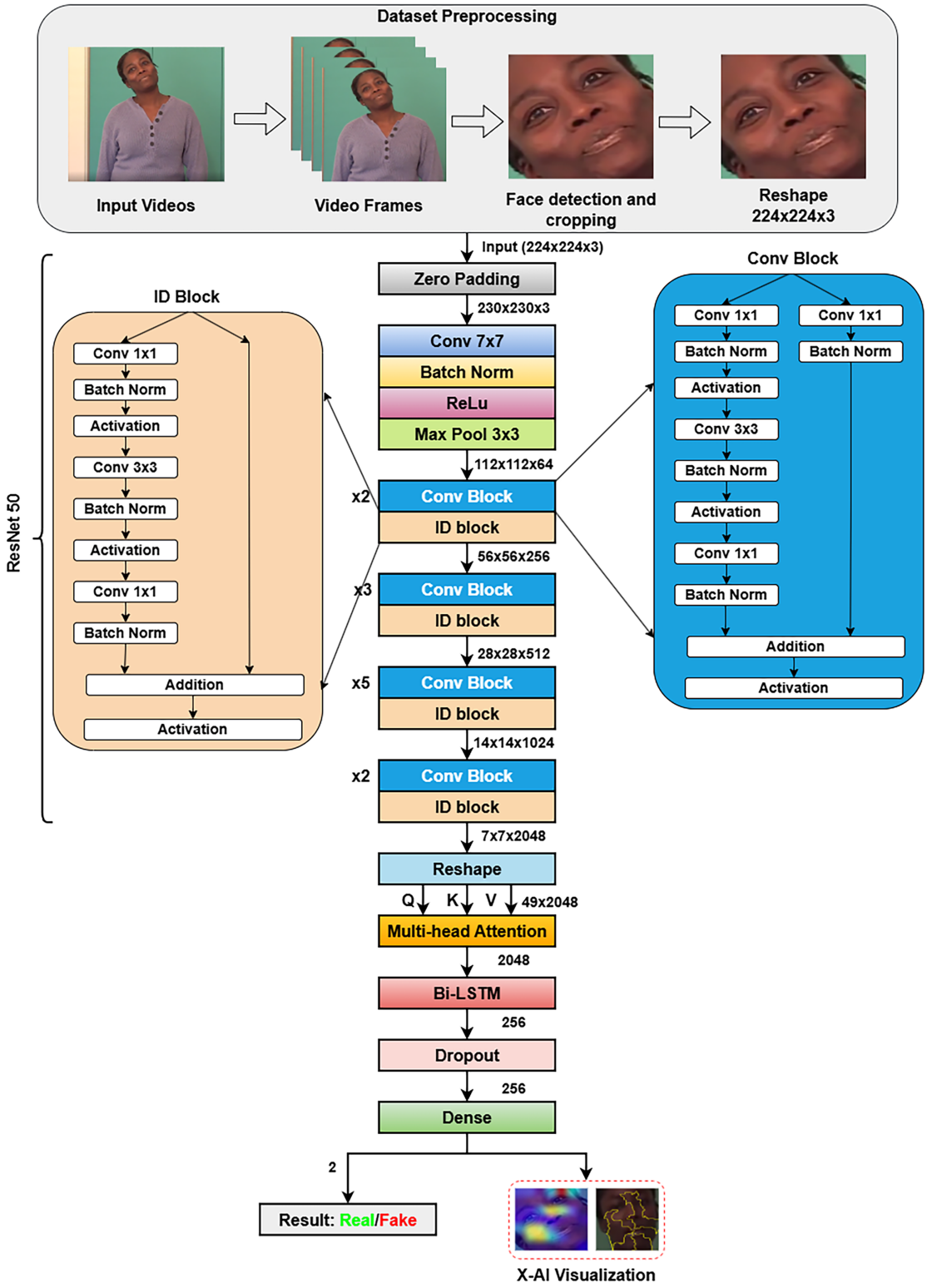


Figure 3: Proposed HCBLNet architecture for deepfake video detection.

Here, the feature map F is reshaped into a sequence of spatial tokens $F_r \in \mathbb{R}^{49\times 2048}$ for attention processing, heads h to process the sequence. For each head i, the input sequence is linearly transformed to obtain query Q_i, key K_i, and value V_i matrices.

$$Q_i = F_r W_i^Q, \quad K_i = F_r W_i^K, \quad V_i = F_r W_i^V$$

where $W^V, W^K, W^Q \in \mathbb{R}^{d_{model}\times d_k}$.

Using the scaled dot-product, the attention scores (A_i) are computed based on the interaction between query and key vectors.

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \tag{2}$$

Each attention head produces its output by a matrix multiplication between the attention scores and the value matrix V_i.

$$O_i = A_i V_i \tag{3}$$

The outputs of the individual attention heads are combined through concatenation, followed by a linear transformation to produce the final result of the multi-head attention layer.

$$O = \text{Concat}(O_1, O_2, \ldots, O_h) W^O \tag{4}$$

where $W^O \in \mathbb{R}^{h\cdot d_v \times d_{model}}$.

The resulting output of the multi-head attention layer preserves the spatial token dimension while incorporating enriched contextual spatial information across the 49 regions of the frame.

3.3.3 Sequence Modeling with Bidirectional LSTM

To model dependencies among the spatially enhanced regions of a frame, a Bidirectional LSTM layer is integrated to capture contextual information by treating the sequence of 49 spatial tokens (produced by the attention module) as an ordered sequence and processing it in both forward and backward directions. We clarify that, in the current formulation, this recurrent module operates over the intra-frame spatial-token sequence rather than over a sequence of video frames; no multi-frame clip is constructed, and HCBLNet does not perform temporal aggregation across frames. BiLSTM layer has 128 units in each direction, capturing dependencies among the spatial token sequence. This bidirectional design enables the model to leverage information from spatial tokens on either side of the sequence ordering, thereby improving its comprehension of the overall frame content. To reduce the risk of overfitting, a dropout layer with a 0.5 rate is employed, which randomly deactivates a portion of the input neurons during training. Furthermore, L2 regularization is applied to the LSTM component to constrain weight updates, promoting better generalization to previously unseen data. The output O from the multi-head attention layer is fed into a Bidirectional LSTM with 128 units per direction. Let H^b and H^f denote the hidden state outputs of the backward and forward LSTM layers, respectively.

For each token position t ($t = 1, \ldots, 49$, indexing spatial tokens within the frame):

$$H_t^f = \text{LSTM}_f(H_{t-1}^f, O_t), \quad H_t^b = \text{LSTM}_b(H_{t+1}^b, O_t) \tag{5}$$

For the forward LSTM (H_t^f):

$$i_t^f = \sigma(W_i^f O_t + U_i^f H_{t-1}^f + b_i^f)$$
$$f_t^f = \sigma(W_f^f O_t + U_f^f H_{t-1}^f + b_f^f)$$
$$o_t^f = \sigma(W_o^f O_t + U_o^f H_{t-1}^f + b_o^f)$$
$$\tilde{C}_t^f = \tanh(W_C^f O_t + U_C^f H_{t-1}^f + b_C^f)$$
$$C_t^f = f_t^f \odot C_{t-1}^f + i_t^f \odot \tilde{C}_t^f$$
$$H_t^f = o_t^f \odot \tanh(C_t^f)$$

For the backward LSTM (H_t^b):

$$i_t^b = \sigma(W_i^b O_t + U_i^b H_{t+1}^b + b_i^b)$$
$$f_t^b = \sigma(W_f^b O_t + U_f^b H_{t+1}^b + b_f^b)$$
$$o_t^b = \sigma(W_o^b O_t + U_o^b H_{t+1}^b + b_o^b)$$
$$\tilde{C}_t^b = \tanh(W_C^b O_t + U_C^b H_{t+1}^b + b_C^b)$$
$$C_t^b = f_t^b \odot C_{t+1}^b + i_t^b \odot \tilde{C}_t^b$$
$$H_t^b = o_t^b \odot \tanh(C_t^b)$$

Hidden states in backward and forward directions are concatenated for each token position t:

$$H_t = \text{Concat}(H_t^f, H_t^b) \tag{6}$$

A dropout layer with a rate of 0.5 and L2 regularization is applied. The hidden states across all 49 token positions are then aggregated (via global average pooling) into a single frame-level representation, which is passed to the final dense layer with softmax activation that classifies the frame as real or fake.

$$y = \text{softmax}(W^H H_t + b^H) \tag{7}$$

To mitigate overfitting, a dropout layer is incorporated following the Bidirectional LSTM layer, which functions by randomly deactivating a portion of the input units during each training iteration. This helps to discourage the proposed model from becoming high dependent on any particular neurons, thereby enhancing its generalization capability. Following the dropout layer, the final dense layer aggregates the learned features. The output layer has 2 units corresponding to the binary classification (real or fake). To further control overfitting, L2 regularization is employed on the final dense layer, discouraging excessively large weight values. This layer utilizes a softmax activation function, enabling the model to classify each frame as real or fake by generating a probability distribution across the two categories. We emphasize that this classification is performed independently for each frame; the reported real/fake decision corresponds to a single input frame.

3.3.4 Interpretability with Grad-CAM and LIME Visualization

To enhance model interpretability and provide visual explanations for classification decisions, we incorporate the Grad-CAM [19] and LIME [51] explainable AI technique. Grad-CAM produces localization maps specific to each class by leveraging the gradients of the target class that propagate into the CNN's last

convolutional layer. This approach helps identify the spatial regions in the input video frames that have the greatest impact on the model's decision-making process, fostering trust and facilitating error analysis.

Let y^c represent the output score corresponding to class c (either real or fake), and let $A^k \in \mathbb{R}^{u \times v}$ denote the k-th feature map obtained from the final convolutional layer of ResNet50. In Grad-CAM, the importance weight α_k^c assigned to each feature map is computed by applying global average pooling to the gradient of the class score y^c with respect to the activations A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{8}$$

where Z represents the total number of pixels contained within the feature map A^k (i.e., $Z = u \times v$). The class-specific localization map $L_{\text{Grad-CAM}}^c \in \mathbb{R}^{u \times v}$ is obtained by computing a weighted sum of the feature maps, followed by applying a ReLU activation to retain only the features that contribute positively to the prediction of the target class:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{9}$$

While Grad-CAM provides class-discriminative spatial localization, we also employ Local Interpretable Model-Agnostic Explanations (LIME) [51] to obtain complementary interpretability at the superpixel level. LIME explains the prediction of a complex model by approximating it locally with an interpretable surrogate model g around an input instance x. Given an input image x, LIME generates a perturbed dataset $\{x_i'\}$ by randomly masking superpixels and obtains corresponding predictions $f(x_i')$ from the trained model. A weighted linear model g is then learned by minimizing:

$$\xi(x) = \arg\min_{g \in G} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{10}$$

where $\mathcal{L}$ measures the fidelity of the surrogate model g in approximating the original model f, π_x defines the proximity measure between perturbed samples and the original input, and $\Omega(g)$ enforces interpretability by penalizing model complexity. The final explanation identifies the most influential superpixels contributing to the classification decision.

The resulting LIME visualization highlights the image regions that most influence the prediction at a local level, complementing Grad-CAM by offering model-agnostic interpretability. Unlike Grad-CAM, which relies on internal gradients of the CNN architecture, LIME provides flexible post-hoc explanations applicable to any classifier. However, due to its superpixel-based perturbation strategy, LIME explanations may appear more fragmented compared to the spatially smooth heatmaps generated by Grad-CAM.

Integrating these XAI techniques into our model not only aids in qualitative analysis but also assists researchers and practitioners in verifying whether the model is focusing on relevant facial regions when making decisions.

3.4 Model Training and Evaluation

The proposed HCBLNet for frame-level deepfake detection was trained independently using three prominent deepfake video datasets, namely FaceForensics++, DFDC, and Celeb-DF. The trained models from these three distinct datasets were then tested on the respective test datasets. The results demonstrate the effectiveness of the proposed method for frame-level deepfake detection on the evaluated benchmarks.

CNNs, particularly architectures like ResNet50, are exceptional feature extractors due to their deep structure and the use of residual connections that help in training very deep networks. ResNet50, with its 50 layers, is capable of capturing intricate details in images, making it highly effective for extracting rich spatial features that can highlight subtle manipulations in deepfake videos. Subsequently, multi-head attention enables the model to simultaneously focus on various features of the input by allocating distinct attention weights to different segments of the data. This is particularly useful for detecting deepfakes, as it enables the model to capture both local and global inconsistencies among the spatial regions of a single frame. Finally, Bidirectional LSTM networks play a vital role in capturing contextual dependencies among the spatially enhanced regions of a frame. BiLSTMs analyze input sequences bidirectionally, enabling the model to learn contextual cues and dependencies across the 49 spatial tokens produced from each frame. This two-way processing enables the model to identify spatial anomalies and inconsistencies within a frame commonly found in deepfake content. In the proposed method, we combine CNN, multi-head attention, and BiLSTM to create a comprehensive deepfake recognition system. By combining the strengths of each component, our system aims to achieve robust detection of manipulated frames by jointly modeling fine-grained spatial features and their contextual dependencies within each frame.

To improve interpretability, we incorporated Grad-CAM into the proposed architecture. Grad-CAM helps visualize which regions of a video frame the model focuses on while making predictions. By computing the gradient of the output class with respect to the final convolutional feature maps in ResNet50, Grad-CAM generates heatmaps that highlight discriminative regions for classification. This visualization technique enhances the transparency of our model by offering insight into its spatial reasoning process and helps validate whether the model is attending to plausible and relevant features when identifying manipulated content.

In this study, we compare the proposed model with some popular transfer-learning-based models and existing work in the literature on this specific application. We fine-tuned common hyperparameters across all models to ensure a fair comparison and obtain the best possible results from each model. Table 2 displays the hyperparameter optimization space for all experimented models. This network optimization approach involved training each type of classifier independently. Subsequently, the model that achieved the highest test accuracy for each deepfake detection task was selected. Following that, we plotted confusion matrices and ROC curves for the classifiers and conducted a 5-fold cross-validation experiment. Finally, we compared the best results obtained by the selected model for detecting deepfake frames with each other and existing literature.

Table 2: List of hyperparameters and their corresponding search space used for fine-tuning the experimented models. Bold values indicate the final selected hyperparameters, including the chosen optimizer.

Criteria	Search Space
Models	[VGG16, VGG19, ResNet50, ResNet152, DenseNet121, DenseNet201, EfficientNet, Multihead CNN-BiLSTM (VGG16), HCBLNet]
Pre-training datasets	[ImageNet-1k]
Epochs	[30, 40, **50**, 60]
Batch sizes	[16, **32**]
Optimizers	[**Adam**, SGD]
Loss functions	[Categorical Cross Entropy]
Learning rates	[0.03, 0.01, **0.001**]

(Continued)

Table 2 (continued)

Criteria	Search Space
l2 regularizer	[**0.01**, 0.001]
NUM heads	[4, **8**, 12] (only for proposed model)
Dropout	[0.1, 0.3, **0.5**]
LSTM units	[64, **128**, 256] (only for proposed model)

All experiments were conducted on the Kaggle platform using 2× Tesla T4 GPUs. The implementation was developed using Python 3.9 with the TensorFlow/Keras framework. Owing to the differences in dataset size and convergence behavior, the training time varied across datasets. On average, training required approximately 20–30 min for the DFDC dataset, 40–50 min for FaceForensics++, and 90–100 min for Celeb-DF. These variations are primarily attributed to the increasing number of training samples and the number of epochs completed before early stopping. No explicit data augmentation techniques were applied during training, as the models were trained directly on preprocessed frames extracted from the videos. Early stopping with a patience of 5 epochs, monitored using the validation loss, was employed to prevent overfitting, and the best-performing model weights were restored automatically. The Adam optimizer with a learning rate of 0.001 was selected as the final optimization strategy, as summarized in Table 2.

Algorithm 1 outlines the architecture and training process of the proposed HCBLNet model to effectively detect deepfake content at the frame level.

Algorithm 1: DeepFake detection using HCBLNet with explainable AI

Input: Video frames $F = \{f_1, f_2, \ldots, f_n\}$, Pre-trained ResNet50, number of heads h, LSTM units u, dropout rate p, L2 regularization rate λ, number of classes c, learning rate α
Output: Predicted class y_i for each frame $f_i \in F$
function DeepFakeDetection(F, $ResNet50$, h, u, p, λ, c, α)
    $resnet\_model \leftarrow$ LoadPretrainedResNet50()
    Freeze layers of $resnet\_model$
    $predictions \leftarrow [\,]$
    for each frame f_i in F **do**
        $R_i \leftarrow$ ExtractFeatures($resnet\_model, f_i$) ▷ $R_i \in \mathbb{R}^{7\times7\times2048}$
        $R_i^{seq} \leftarrow$ Reshape($R_i, (49, 2048)$) ▷ 49 intra-frame spatial tokens
        $A_i \leftarrow$ MultiHeadAttention($R_i^{seq}, h, d\_model = 2048$) ▷ Spatial self-attention among the 49 tokens of frame f_i
        $L_i \leftarrow$ BidirectionalLSTM($A_i,\ units = u,\ dropout\_rate = p,\ l2\_rate = \lambda$) ▷ Models dependencies across the 49 spatial tokens of frame f_i
        $V_i \leftarrow$ GlobalAveragePooling(L_i) ▷ Aggregate spatial tokens into a single frame-level representation
        $output_i \leftarrow$ DenseLayer($V_i,\ units = c,\ activation =$ 'softmax', $l2\_rate = \lambda$)
        Append $output_i$ to $predictions$
    end for
    $model \leftarrow$ CompileModel(Adam(α), categorical_crossentropy, accuracy)
    TrainModel($model, X_{train}, y_{train}, X_{val}, y_{val}$)

(Continued)

Algorithm 1 (continued)

$y \leftarrow predictions$ ▷ Per-frame predicted classes for all frames in F
$last\_conv\_layer \leftarrow$ GetLastConvLayer($resnet\_model$)
$gradcam\_map \leftarrow$ GenerateGradCAM($resnet\_model, last\_conv\_layer, model, F$)
Overlay $gradcam\_map$ on the input frames
$lime\_explanation \leftarrow$ GenerateLIME($model, F$)
Visualize $lime\_explanation$ using superpixel perturbation
return y
end function

4 Result and Analysis

In this section, we present a comprehensive evaluation of the proposed HCBLNet for deepfake video detection. We first analyze the performance of our model on three widely used datasets, followed by comparisons with existing transfer learning models and prior research. Both holdout testing and 5-fold cross-validation experiments are conducted to assess the robustness, generalizability, and reliability of the proposed approach. The evaluation focuses on four key metrics: accuracy, precision, recall, and F1-score, which together provide a holistic view of the model's detection capabilities. Additionally, we perform an ablation study to quantify the contribution of each major component, including the ResNet50 backbone, BiLSTM, and multi-head attention, on the overall performance. Finally, Grad-CAM and LIME visualizations are used to interpret the model's decision-making process, highlighting the facial regions most relevant for deepfake detection and demonstrating the model's ability to focus on semantically meaningful and manipulation-sensitive areas.

4.1 Performance on Individual Datasets

To better understand the effectiveness of the proposed HCBLNet, we first evaluate its performance separately on the three datasets: DFDC, FaceForensics++, and Celeb-DF. Table 3 summarizes the model's performance in terms of accuracy, precision, recall, and F1-score on each dataset.

Table 3: Proposed HCBLNet performance on individual datasets.

Dataset	Accuracy	Precision	Recall	F1-Score
DFDC	97.91%	100%	96.34%	98.13%
FaceForensics++	95.40%	93.71%	97.68%	95.65%
Celeb-DF	95.97%	94.74%	98.57%	96.62%

On the DFDC dataset, HCBLNet achieves an accuracy of 97.91%, with perfect precision of 100%, a recall of 96.34%, and an F1-score of 98.13%. These results indicate that all detected deepfake videos were true positives, while the model was able to identify nearly all actual deepfake instances. For FaceForensics++, the model obtains an accuracy of 95.40%, precision of 93.71%, recall of 97.68%, and F1-score of 95.65%. The slightly lower precision compared to DFDC is due to a few false positives, but the high recall demonstrates the model's ability to detect most manipulated videos in the dataset. On the Celeb-DF dataset, HCBLNet achieves an accuracy of 95.97%, precision of 94.74%, recall of 98.57%, and F1-score of 96.62%. The high recall highlights the model's robustness in identifying almost all deepfake instances, while the balanced precision ensures minimal misclassification of real videos.

These results collectively demonstrate that the proposed HCBLNet is highly effective across datasets of varying sizes and qualities. The model consistently achieves strong performance metrics, confirming its ability to generalize to different types of deepfake manipulations while maintaining low false positive rates.

4.2 Ablation Study

To investigate the individual contribution of each component in the proposed HCBLNet architecture, we conducted a systematic ablation study by progressively adding modules and evaluating their impact on performance. This study examines how each component, including ResNet50, BiLSTM, and multi-head attention, contributes to the final detection performance. All ablation experiments were conducted using the same datasets, preprocessing strategy, training protocol, and evaluation metrics to ensure a fair comparison. Performance is reported using accuracy, precision, recall, and F1-score on the DFDC, FaceForensics++, and Celeb-DF datasets.

First, we considered the base spatial model, where only ResNet50 was used as a frame-level feature extractor and classifier without any spatial-sequence or attention-based modeling. This configuration relies purely on raw spatial cues from individual frames without modeling relationships among the spatial regions within each frame. As expected, the performance was comparatively lower across all datasets. On the DFDC dataset, the ResNet50-only model achieved an accuracy of 96.12%, precision of 98.26%, recall of 94.41%, and F1-score of 96.30%. On FaceForensics++, it achieved 93.25% accuracy, 91.91% precision, 95.59% recall, and 93.71% F1-score. Similarly, on Celeb-DF, the model achieved an accuracy of 94.21%, precision of 94.51%, recall of 95.65%, and F1-score of 95.07%. These results indicate that spatial features alone, extracted independently without modeling dependencies among spatial regions within a frame, are insufficient to fully capture the subtle manipulation artifacts present in deepfake videos.

Next, we evaluated the effect of combining ResNet50 with BiLSTM, forming a CNN-BiLSTM architecture without attention. This configuration introduces sequential modeling over the spatial-token sequence derived from each frame's 7×7 feature grid, enabling the model to capture dependencies among different spatial regions within the same frame, such as inconsistencies between facial regions that may indicate manipulation artifacts. We clarify that this sequential modeling operates within a single frame and does not involve any cross-frame or video-level processing. The inclusion of BiLSTM led to a noticeable improvement in recall and overall performance across all datasets. On the DFDC dataset, the CNN-BiLSTM model achieved an accuracy of 97.10%, precision of 98.85%, recall of 95.62%, and F1-score of 97.22%. On FaceForensics++, accuracy increased to 94.35%, precision to 92.80%, recall to 96.74%, and F1-score to 94.73%. For Celeb-DF, the model achieved 95.12% accuracy, 94.03% precision, 97.12% recall, and 95.55% F1-score. These results confirm that modeling dependencies among spatial regions within a frame significantly enhances deepfake detection performance, particularly in terms of recall.

Finally, we evaluated the full HCBLNet model, which integrates ResNet50, BiLSTM, and multi-head attention. This configuration enables the model to simultaneously capture spatial features, intra-frame spatial-token dependencies, and attention-weighted discriminative regions within each frame. The full model achieved the best performance across all datasets and metrics. On the DFDC dataset, HCBLNet achieved an accuracy of 97.91%, precision of 100%, recall of 96.34%, and F1-score of 98.13%. On FaceForensics++, the model achieved 95.40% accuracy, 93.71% precision, 97.68% recall, and 95.65% F1-score. Similarly, on the Celeb-DF dataset, HCBLNet achieved an accuracy of 95.97%, precision of 94.74%, recall of 98.57%, and F1-score of 96.62%.

The progressive improvement observed from using a single component to combining multiple components clearly demonstrates that each module contributes meaningfully to the final performance. While ResNet50 provides strong spatial representations, the addition of BiLSTM enhances spatial coherence

modeling by capturing dependencies among the 49 spatial tokens within each frame, and the incorporation of multi-head attention further refines feature discrimination by focusing on manipulation-sensitive regions within the frame. As summarized in Table 4, the incremental integration of these components consistently improves accuracy, precision, recall, and F1-score across all three datasets. The results confirm that the best performance is achieved only when all three components are jointly integrated, thereby justifying the architectural design choices of the proposed HCBLNet.

Table 4: Ablation study results on DFDC, FaceForensics++, and Celeb-DF datasets. Bold value represent the best result.

Model Configuration	Dataset	Accuracy	Precision	Recall	F1-Score
ResNet50 only	DFDC	96.12%	98.26%	94.41%	96.30%
	FaceForensics++	93.25%	91.91%	95.59%	93.71%
	Celeb-DF	94.21%	94.51%	95.65%	95.07%
ResNet50 + BiLSTM	DFDC	97.10%	98.85%	95.62%	97.22%
	FaceForensics++	94.35%	92.80%	96.74%	94.73%
	Celeb-DF	95.12%	94.03%	97.12%	95.55%
HCBLNet	DFDC	**97.91%**	**100%**	**96.34%**	**98.13%**
	FaceForensics++	**95.40%**	**93.71%**	**97.68%**	**95.65%**
	Celeb-DF	**95.97%**	**94.74%**	**98.57%**	**96.62%**

To further justify the architectural design choices of HCBLNet, we conducted a hyperparameter sensitivity analysis by varying key model parameters while keeping the remaining settings unchanged. It is worth noting that the full hyperparameter grid comprising all combinations of epochs, batch size, optimizer, learning rate, L2 regularizer, number of attention heads, dropout rate, and BiLSTM units, as detailed in Table 2, yields 1944 possible configurations. Exhaustively evaluating this entire search space is computationally impractical; therefore, we adopted a random search strategy, sampling 30–40 representative combinations for each dataset. The top-5 performing configurations obtained through this procedure are summarized in Table 5. Specifically, we evaluated the impact of different optimizers, numbers of attention heads, BiLSTM units, and dropout rates on the performance of the proposed model across the DFDC, FaceForensics++, and Celeb-DF datasets. The results demonstrate that the Adam optimizer consistently outperformed SGD across all three datasets. Using Adam resulted in the highest accuracy, recall, and F1-score, achieving 97.91%, 95.40%, and 95.97% accuracy on DFDC, FaceForensics++, and Celeb-DF, respectively, along with competitive precision on all three datasets. This indicates that Adam provides more effective optimization for the proposed architecture by enabling faster and more stable convergence during training. We also investigated the influence of the number of attention heads. As shown in Table 5, configurations using 8 attention heads consistently achieved higher or comparable performance relative to configurations using 4 and 12 heads across all three datasets. This suggests that 8 heads provide an effective balance between feature diversity and model complexity, allowing the model to better capture discriminative manipulation patterns in deepfake videos without introducing unnecessary redundancy. Similarly, the effect of varying the number of BiLSTM units was analyzed. The results indicate that 128 BiLSTM units consistently achieved the highest performance compared to 64 and 256 units. While 64 units may not provide sufficient representational capacity to model dependencies among the spatial tokens within a frame, increasing the number to 256 introduces additional complexity without corresponding performance gains. Therefore, 128 units were selected as the optimal configuration for capturing discriminative spatial-token relationships within individual frames of manipulated videos. Finally, we evaluated the impact of different dropout rates.

The results show that a dropout rate of 0.5 yielded the best generalization performance across all datasets. Lower dropout rates of 0.1 and 0.3 resulted in slightly lower performance, indicating a higher tendency toward overfitting. The selected dropout rate effectively regularizes the model while preserving its ability to learn discriminative spatial representations within each frame. Overall, the hyperparameter sensitivity analysis presented in Table 5 validates the design choices adopted in HCBLNet. The selected configuration consisting of the Adam optimizer, 8 attention heads, 128 BiLSTM units, and a dropout rate of 0.5 consistently achieved the highest or near-highest performance across all evaluation metrics and datasets, thereby supporting the effectiveness and robustness of the proposed architecture.

Table 5: Hyperparameter sensitivity analysis of HCBLNet on DFDC, FaceForensics++, and Celeb-DF dts. Bold values indicate the final selected hyperparaatasemeter configuration and best result.

Epochs	Batch Size	Optimizer	Learning Rate	L2 Reg.	Heads	Dropout	BiLSTM	Accuracy	Precision	Recall	F1 Score
DFDC Dataset											
50	**32**	**Adam**	**0.001**	**0.01**	**8**	**0.5**	**128**	**97.91%**	**100.00%**	96.34%	**98.13%**
50	16	SGD	0.010	0.01	4	0.1	128	96.71%	96.74%	**96.72%**	96.72%
40	16	Adam	0.010	0.01	8	0.3	64	96.42%	96.61%	96.42%	96.43%
50	16	SGD	0.001	0.01	4	0.1	128	96.42%	96.55%	96.42%	96.43%
60	16	SGD	0.030	0.01	4	0.5	256	96.42%	96.46%	96.42%	96.42%
FaceForensics++ Dataset											
50	**32**	**Adam**	**0.001**	**0.01**	**8**	**0.5**	**128**	**95.40%**	93.71%	**97.68%**	**95.65%**
50	16	Adam	0.010	0.01	4	0.1	64	95.02%	**95.02%**	95.02%	95.01%
40	32	SGD	0.030	0.01	8	0.5	64	94.63%	94.63%	94.63%	94.63%
60	16	Adam	0.010	0.01	4	0.5	64	94.49%	94.51%	94.50%	94.50%
40	16	SGD	0.010	0.01	8	0.3	64	94.41%	94.46%	94.41%	94.41%
Celeb-DF Dataset											
50	**32**	**Adam**	**0.001**	**0.010**	**8**	**0.5**	**128**	**95.97%**	94.74%	**98.57%**	**96.62%**
40	32	SGD	0.030	0.010	8	0.5	64	95.05%	**95.06%**	95.05%	95.05%
40	16	SGD	0.030	0.001	4	0.1	256	94.90%	94.92%	94.90%	94.90%
40	16	SGD	0.010	0.010	8	0.3	64	94.85%	94.85%	94.85%	94.85%
60	16	Adam	0.010	0.010	4	0.3	128	94.80%	94.80%	94.80%	94.80%

4.3 Cross-Validation and Cross-Dataset Performance Analysis

To further assess the stability and reliability of the proposed HCBLNet architecture, a 5-fold cross-validation was performed using the Celeb-DF, DFDC, and FaceForensics++ datasets. This validation technique was employed to confirm that the model's performance remains consistent across different subsets of the data, ensuring that the results are not biased by a particular train-test split. Such an approach offers a more thorough evaluation of the model's generalization ability. A summary of the cross-validation outcomes is presented in Table 6. Specifically, for the DFDC dataset, the model achieved an average accuracy of 97.36% with a standard deviation of 1.6%, reflecting stable and high performance across all folds. On the FaceForensics++ dataset, the average accuracy was 94.98% with a standard deviation of 1.14%, demonstrating reliable performance with accuracy ranging from 92.96% to 96.09%. On the Celeb-DF dataset, the model attained an average accuracy of 95.17% with a standard deviation of 0.85%, indicating consistent performance across folds, with accuracy values ranging from 93.85% to 96.23%. These cross-validation results confirm the stability, reliability, and generalizability of our model across diverse datasets, reinforcing its potential for effective deepfake identification.

Table 6: 5-fold cross-validation accuracy for the DFDC, FaceForensics++, and Celeb-DF datasets.

Metrics	DFDC	FaceForensics++	Celeb-DF
Fold 1	96.52%	96.09%	93.85%
Fold 2	98.80%	94.58%	94.84%
Fold 3	99.70%	95.33%	95.97%
Fold 4	95.52%	92.96%	96.23%
Fold 5	96.27%	95.92%	94.97%
Average ± Standard Deviation	97.36 ± 1.6%	94.98 ± 1.14%	95.17 ± 0.85%

Beyond standard cross-validation, we evaluated the cross-dataset generalization ability of HCBLNet by training the model on one dataset and testing it on the other two unseen datasets. This assessment provides insights into the model's robustness when encountering unseen video distributions and manipulation types. The results, summarized in Table 7, indicate that the proposed model maintains strong performance even under cross-dataset conditions, although a slight reduction in metrics is observed compared to within-dataset evaluation, which is expected due to domain differences.

Table 7: Cross-dataset evaluation of HCBLNet: Model trained on one dataset and tested on others.

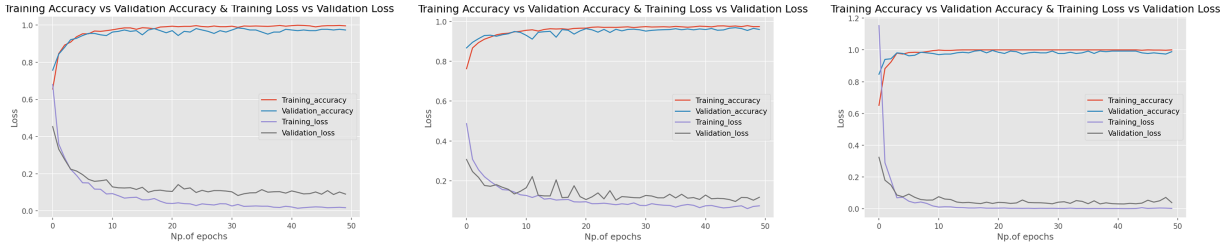
Training Dataset	Testing Dataset	Accuracy	Precision	Recall	F1-Score
DFDC	FaceForensics++	93.84%	92.15%	95.22%	93.66%
	Celeb-DF	94.15%	93.04%	96.11%	94.55%
FaceForensics++	DFDC	95.01%	96.10%	94.25%	95.17%
	Celeb-DF	94.72%	93.56%	95.89%	94.71%
Celeb-DF	DFDC	94.26%	95.12%	93.10%	94.10%
	FaceForensics++	93.98%	92.44%	95.50%	93.95%

When trained on the DFDC dataset, HCBLNet achieved an accuracy of 93.84% on FaceForensics++ and 94.15% on Celeb-DF, demonstrating that the model generalizes well to datasets with different manipulation characteristics and video qualities. Similarly, when trained on FaceForensics++, the model attained an accuracy of 95.01% on DFDC and 94.72% on Celeb-DF, indicating that training on a larger and more diverse dataset helps the model adapt effectively to other datasets. Training on Celeb-DF resulted in an accuracy of 94.26% on DFDC and 93.98% on FaceForensics++, confirming that even when exposed to dataset-specific manipulations, HCBLNet maintains strong detection capability across domains. Overall, these cross-dataset results underscore the robust generalizability of the proposed HCBLNet, showing that it can effectively detect deepfakes even when applied to datasets that were not seen during training.

4.4 Training Behavior and Convergence Analysis

Fig. 4 illustrates the training curves for the Multi-Head Attention Based CNN-BiLSTM model trained on three distinct datasets: DFDC, FaceForensics++, and Celeb-DF during holdout training and validation. The training process involved multiple epochs, with the number of epochs set at 50. We selected this value after experimenting with different epoch numbers, including 30, 40, 50, and 60, to determine the optimal training duration with greater results. Fig. 4 demonstrates that for all datasets, the accuracies consistently increased while the losses decreased over the epochs. This trend indicates that the networks effectively

learned from the training data and generalized well to validation data, suggesting that the models were well-trained without exhibiting signs of overfitting or underfitting.

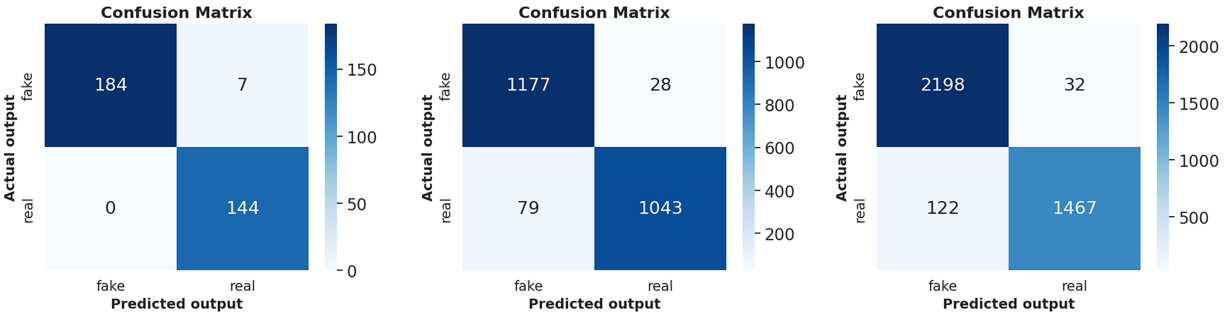


(a) DFDC **(b)** FaceForensics++ **(c)** Celeb-DF

Figure 4: Training curves of the proposed HCBLNet on different datasets.

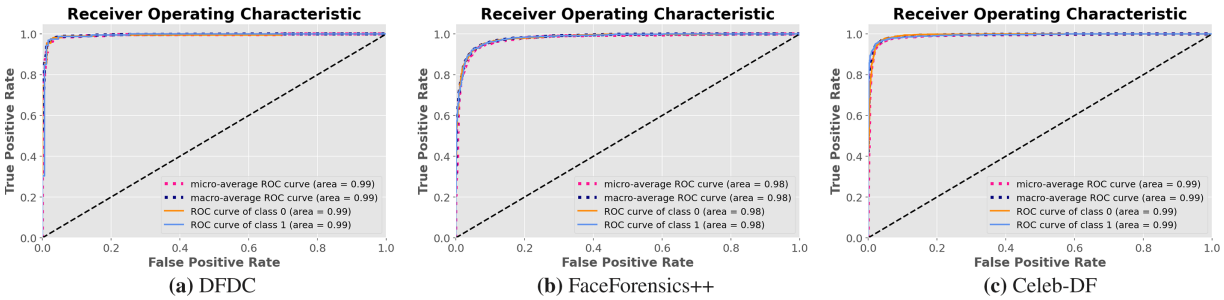
4.5 Confusion Matrix and ROC Curve Analysis

Figs. 5 and 6 display the confusion matrices and ROC curves of the classifiers. The macro and micro AUC values of classifiers trained on three distinct datasets exceeded 0.98, indicating excellent discriminative ability. The confusion matrices show minimal false positives and false negatives across all datasets. Notably, on the DFDC dataset, the proposed model achieved 0 false positives, highlighting its exceptional performance in accurately identifying deepfake videos. Consequently, the precision and recall values were significantly high on all three datasets. The precision rates were 100%, 93.71%, and 94.74% for the three distinct datasets, while the recall rates were 96.34%, 97.68%, and 98.57%, respectively. This low error rate and high accuracy across all datasets underscore the robustness and reliability of the proposed Multi-Head Attention-Based CNN-BiLSTM model for deepfake detection tasks.



(a) DFDC **(b)** FaceForensics++ **(c)** Celeb-DF

Figure 5: Confusion matrices of the proposed HCBLNet across the datasets.



(a) DFDC **(b)** FaceForensics++ **(c)** Celeb-DF

Figure 6: ROC curves of the proposed HCBLNet across all datasets.

4.6 Model Interpretability Analysis Using X-AI

Explainable Artificial Intelligence (XAI) has become increasingly important in deepfake detection because deep learning models often operate as black-box systems, making it difficult to understand the reasoning behind their predictions. In applications such as digital forensics, journalism, cybersecurity, and content verification, understanding why a model classifies a video as real or fake is essential for building trust and supporting decision-making. XAI techniques provide insights into the regions and features that contribute most to model predictions, allowing researchers to verify whether the model focuses on meaningful manipulation artifacts rather than irrelevant visual patterns. Therefore, incorporating XAI methods improves both the transparency and reliability of deepfake detection systems. To further interpret the decision-making process of our model and improve its explainability, we applied Grad-CAM and LIME to visualize the regions that contributed most to the model's deepfake predictions. Fig. 7 presents representative visual explanations generated by the proposed model. Grad-CAM produces class-discriminative heatmaps by leveraging gradient information from the final convolutional layers of the ResNet50 backbone. As shown in Fig. 7, Grad-CAM consistently highlights facial regions commonly affected by deepfake manipulations, including the eyes, nose, mouth, and facial boundaries. The generated heatmaps are spatially coherent and focus on semantically meaningful facial regions rather than irrelevant background areas. This behavior suggests that the model relies on manipulation-sensitive features when making predictions.

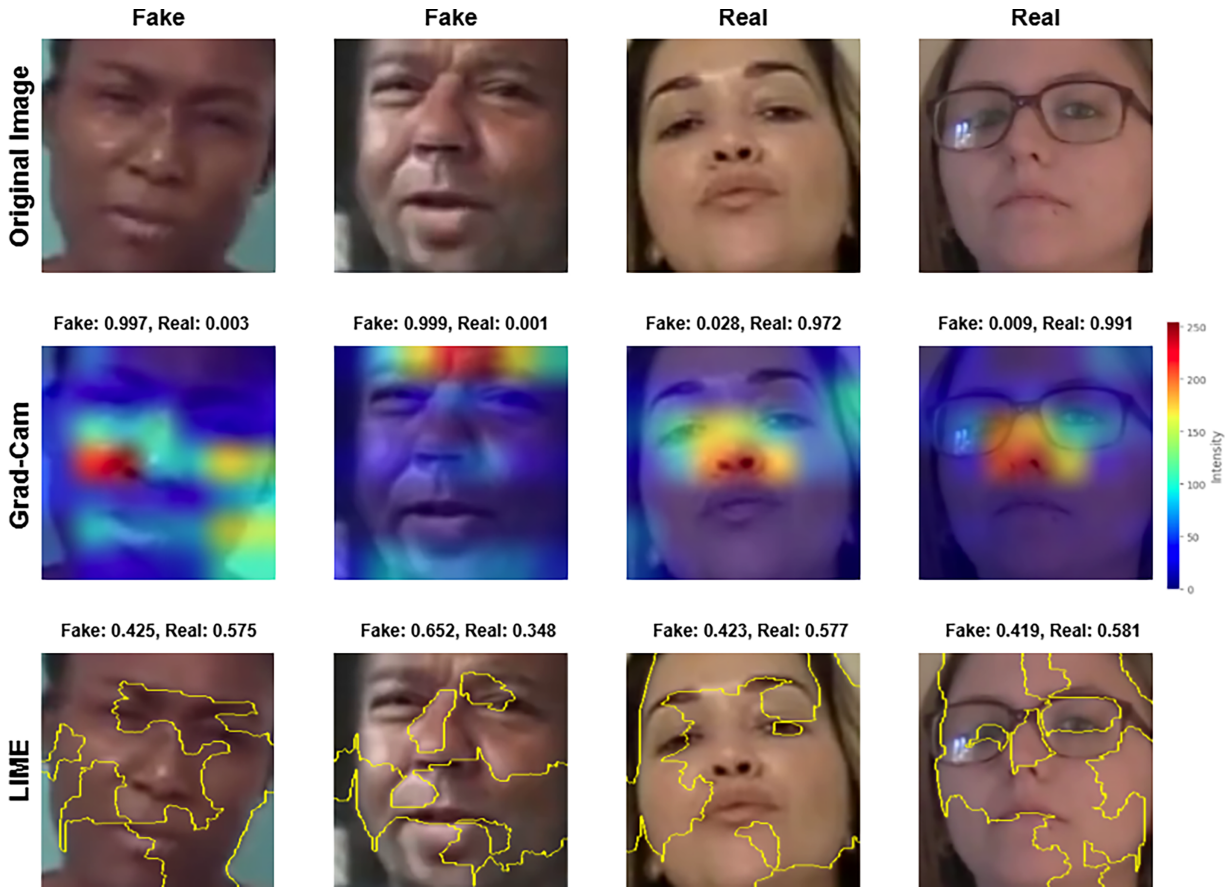


Figure 7: Grad-CAM and LIME visual explanations on real and deepfake frames, highlighting model attention and influential regions.

To provide a broader explainability analysis, we also explored LIME as an alternative XAI technique. LIME explains predictions by perturbing image segments and estimating their contribution to the final classification outcome. While LIME is useful for identifying influential image regions, the resulting explanations are often fragmented due to the superpixel-based segmentation process. The visualizations produced by both techniques indicate that the proposed HCBLNet model focuses on facial areas that are frequently manipulated during the deepfake generation process. This observation supports the validity of the learned feature representations and demonstrates that the model bases its decisions on relevant visual evidence. Overall, the XAI analysis enhances the interpretability of HCBLNet and provides additional confidence in its effectiveness for real-world deepfake detection applications.

4.7 Comparison with Transfer Learning Models

In this subsection, we compare the performance of the proposed HCBLNet with several widely used transfer learning models for deepfake detection. The goal is to evaluate how our HCBLNet architecture based on ResNet50 performs relative to popular backbone networks such as VGG, ResNet, DenseNet, and EfficientNet. We fine-tuned the common hyperparameters of all models to ensure a fair comparison and to obtain optimal results. Table 8 presents a comprehensive comparison between our proposed HCBLNet model based on ResNet50 and various transfer learning models on the Dataset DFDC. The evaluation of each model is conducted using four metrics: accuracy, precision, recall, and the F1-score. Among the transfer learning models, EfficientNet recorded the highest accuracy at 97.18%, while ResNet152 closely followed with an accuracy of 97.01%. These models also demonstrated impressive precision, recall, and F1-scores, indicating their effectiveness in detecting deepfake videos. Our proposed HCBLNet model based on ResNet50 outperformed all other transfer learning models, achieving an accuracy of 97.91%. Notably, it achieved a perfect precision of 100%, indicating that all identified deepfake instances were indeed true positives. Although the recall for our model was slightly lower compared to ResNet152, it still exhibited strong performance with a recall of 96.34% and an F1-score of 98.13%.

Table 8: Comparison with transfer learning models and our proposed HCBLNet on the DFDC Dataset. Bold value represent the best result.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	93.73%	94.30%	94.79%	94.55%
VGG19	95.82%	97.31%	95.26%	96.28%
ResNet50	96.12%	98.26%	94.41%	96.30%
ResNet152	97.01%	97.67%	96.55%	97.11%
DenseNet121	93.13%	94.80%	92.13%	93.45%
DenseNet201	96.72%	97.79%	96.20%	96.99%
EfficientNet	97.18%	96.45%	97.95%	97.20%
Multihead CNN-BiLSTM (VGG16)	96.42%	95.65%	97.78%	96.70%
Proposed HCBLNet	**97.91%**	**100%**	**96.34%**	**98.13%**

Moving on to Table 9, we continue our evaluation by comparing our proposed HCBLNet model with various transfer learning models on the Dataset FaceForensics++. Our proposed model achieved an accuracy of 95.40%, indicating its effectiveness in correctly classifying videos as real or deepfake. The precision of 93.71% signifies the proportion of correctly identified deepfake instances among all detected instances, ensuring minimal false positives. Moreover, the recall of 97.68% highlights the model's ability to

correctly identify most of the actual deepfake instances present in the dataset. The F1-score of 95.65% further emphasizes the balanced performance of the model in terms of both precision and recall. In comparison with other transfer learning models evaluated on the FaceForensics++ dataset, ResNet152 achieved the highest accuracy among traditional models at 94.62%. The proposed model surpasses this performance, achieving higher accuracy and F1-score. EfficientNet also demonstrated competitive results with an accuracy of 94.89%; nevertheless, the proposed model outperformed it across accuracy, precision, recall, and F1-score, indicating superior discrimination between real and deepfake videos.

Table 9: Comparison with transfer learning models and our proposed HCBLNet on the FaceForensics++ Dataset. Bold value represent the best result.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	93.77%	93.42%	94.44%	93.93%
VGG19	93.30%	92.40%	95.11%	93.73%
ResNet50	93.25%	91.91%	95.59%	93.71%
ResNet152	94.62%	93.71%	95.68%	94.68%
DenseNet121	89.69%	86.09%	95.31%	90.47%
DenseNet201	92.42%	96.52%	88.82%	92.51%
EfficientNet	94.89%	93.50%	96.96%	95.20%
Multihead CNN-BiLSTM (VGG16)	94.59%	92.56%	97.44%	94.94%
Proposed HCBLNet	**95.40%**	**93.71%**	**97.68%**	**95.65%**

Further evaluation on the Celeb-DF dataset, as presented in Table 10 reveals that our proposed HCBLNet model continues to demonstrate superior performance. Our model achieved an accuracy of 95.97%, surpassing all other models in the comparison. Specifically, our model's precision of 94.74% ensures that most detected deepfake instances are correctly identified, while the recall of 98.57% shows its capability to identify almost all actual deepfake instances. The F1-score of 96.62% indicates a well-balanced performance in terms of both precision and recall. Among the traditional models, EfficientNet performed well with an accuracy of 95.08%, closely followed by DenseNet201 with 94.95%. The proposed model outperforms these approaches across all evaluation metrics, demonstrating greater robustness and generalizability in deepfake detection scenarios.

Table 10: Comparison with transfer learning models and our proposed HCBLNet on the Celeb-DF dataset. Bold value represent the best result.

Model	Accuracy	Precision	Recall	F1-Score
VGG16	93.24%	93.21%	95.38%	94.28%
VGG19	93.95%	93.74%	96.05%	94.88%
ResNet50	94.21%	94.51%	95.65%	95.07%
ResNet152	94.55%	94.77%	95.96%	95.37%
DenseNet121	94.63%	94.98%	95.87%	95.43%
DenseNet201	94.95%	95.29%	96.10%	95.69%
EfficientNet	95.08%	95.38%	96.23%	95.80%
Multihead CNN-BiLSTM (VGG16)	93.82%	92.19%	97.94%	94.98%
Proposed HCBLNet	**95.97%**	**94.74%**	**98.57%**	**96.62%**

4.8 Comparison with Existing Deepfake Detection Methods

The performance of the proposed HCBLNet model is evaluated using both holdout validation and 5-fold cross-validation across three widely used deepfake datasets, namely DFDC, FaceForensics++, and Celeb-DF. The quantitative results are summarized in Tables 3, 4 and 6–10, while Figs. 4–7 provide further insights into the training dynamics, classification behavior, and model interpretability. Across all evaluation settings, HCBLNet consistently achieves high accuracy, precision, recall, and F1-score, demonstrating strong robustness and reliable generalization across diverse datasets.

To further validate the effectiveness of the proposed approach, we compare HCBLNet with a broad range of recent deepfake detection methods reported in the literature, as summarized in Table 11. Recent studies published between 2021 and 2026 have explored advanced architectures, including temporal modeling and contrastive learning strategies. For instance, Das and Naskar [52] proposed a Bi-GRU-based temporal framework and reported strong performance on DFDC and FakeAVCeleb. Similarly, Zhang et al. [53] introduced a dual-stream contrastive learning approach evaluated across multiple datasets, achieving competitive results on FaceForensics++ but exhibiting reduced accuracy on DFDC and Celeb-DF. More recently, FRENet [54] and HSFF-Net [55] leveraged frequency-based and hierarchical spatial–frequency features, respectively, yet their performance remains limited on challenging datasets such as DFDC and Celeb-DF.

Table 11: Comparison with existing deepfake detection research.

Ref.	Year	Model	No of Dataset	Accuracy
[52]	2026	Bi-GRU	2	DFDC: 95.45% FakeAVCeleb: 99.06%
[53]	2026	Contrastive Learning	4	FaceForensics++: 97.14% DFD: 95.12% Celeb-DF-V2: 86.33% DFDC: 77.93%
[54]	2025	FRENet	4	DFDC: 76.96% Celeb-DF-v1: 88.48% Celeb-DF-v2: 80.40% FaceShifter: 71.14%
[55]	2025	HSFF-Net	4	DFDC: 74.03% FaceForensics++: 75.84% Celeb-DF: 76.24% DFD: 90.86%
[34]	2024	VGG16+LSTM	1	Celeb-DF: 96.25%
[35]	2024	DINOV2+xDNN	1	Celeb-DF: 95.1%
[36]	2024	ResNet-Swish-BiLSTM	2	DFDC: 78.33% Faceforensics++: 95.23%
[37]	2024	Convolution+Self-Attention Network	2	Faceforensics++: 91.69% Celeb-DF: 97.81%

(Continued)

Table 11 (continued)

Ref.	Year	Model	No of Dataset	Accuracy
[38]	2024	Video Vision Transformer (ViViT)	3	CelebDF: 87.17% DFDC 88.03% FaceForencsics++ 90.27%
[39]	2024	DoubleStream+CViT Video Stream Detection Models	3	LRS2: 95.27% FaceForencsics++: 95.10% DFDC: 94.53%
[43]	2023	RIDNet+Multi-Head Relative-Interaction Method	4	FaceForencsics++: 84.36% DFDC: 59.87% Celeb-DF: 70.10% DF-1.0: 67.49%
[41]	2023	Action Unit+Multi-Dependency Graph Module	1	FaceForencsics++: 92.89%
[46]	2023	Deep Convolutional Transformer	4	FaceForencsics++: 92.11% DFDC: 65.76% Celeb-DF: 63.27% DF-1.0: 62.46%
[44]	2022	ResNet+Vision Transformer	5	Celeb-DF: 80.48% Celeb-DFv2: 87.23% FaceForensics++: 75.62% FF-Deepfake Detection: 78.45% DFDC2: 84.55%
[42]	2022	InceptionResNet-BiLSTM	2	FaceForensics++: 93.39% FakeAVCeleb: 77.2%
[40]	2021	CNN+Vision Transformer (ViT)	1	DFDC: 91.50%
[45]	2021	Temporal ID Network+3DMM Generative Network	2	DFDC: 80.4% Celeb-DF: 71.6%
Proposed HCBLNet	-	Multi-Head Attention Based CNN-BiLSTM	3	DFDC: 97.91% Faceforensics++: 95.40% Celeb-DF: 95.97%

Compared to these recent approaches, HCBLNet achieves superior and more consistent performance across all three benchmark datasets. Specifically, our model attains an accuracy of 97.91% on DFDC, outperforming recent temporal and contrastive learning-based methods that report accuracies below 96% on the same dataset [52,53]. On the FaceForensics++ dataset, HCBLNet achieves an accuracy of 95.40%, exceeding the performance of convolutional self-attention networks [37] and transformer-based models such

as ViViT [38]. Similarly, on the Celeb-DF dataset, our method records an accuracy of 95.97%, surpassing both frequency-based networks [54] and vision transformer variants [44].

In addition to accuracy, HCBLNet demonstrates consistently higher precision, recall, and F1-score compared to existing CNN-based and hybrid models. For example, when compared with EfficientNet-based approaches, which report competitive accuracy on Celeb-DF, the proposed model achieves higher recall and F1-score, indicating a stronger ability to correctly identify manipulated content while maintaining balanced classification performance. This improvement can be attributed to the integration of multi-head attention with CNN-based spatial feature extraction and BiLSTM-based sequence modeling over the intra-frame spatial-token representation, enabling the network to capture both fine-grained spatial artifacts and contextual dependencies among spatial regions within each frame. We note that this sequence modeling operates within individual frames and does not involve temporal aggregation across video frames, consistent with the frame-level classification design of HCBLNet described in Section 3.3. The findings from our holdout and 5-fold cross-validation experiment are presented in Tables 3 and 6, and Figs. 4–6. Additional interpretability results are shown in Fig. 7. These results clearly indicate that the proposed multi-head attention-based CNN-BiLSTM model achieves sufficient accuracy for deepfake video detection on the three most popular deepfake video datasets. These comprehensive evaluations demonstrate the robustness and effectiveness of our model across different datasets. The consistent performance across both holdout and cross-validation experiments underscores the model's reliability and its potential for real-world applications in combating the proliferation of deepfake content.

5 Discussion

The primary objective of this study was to evaluate the performance of the proposed HCBLNet for deepfake video detection and compare it with various popular transfer-learning-based models and existing works in the literature. We meticulously fine-tuned common hyperparameters for all models to ensure a fair comparison and achieve the best possible results. This involved optimizing parameters such as the number of epochs, batch sizes, optimizers, learning rates, and the inclusion of L2 regularization, particularly for our proposed model. The results demonstrated that our approach of independent training and selection of the best model based on test accuracy was effective in achieving superior performance.

The proposed model consistently outperformed traditional transfer learning models such as VGG16, VGG19, ResNet50, ResNet152, DenseNet121, DenseNet201, and EfficientNet across three prominent datasets: DFDC, FaceForensics++, and Celeb-DF. Notably, our model achieved an accuracy of 97.91% on the DFDC dataset, which is significantly higher than the best-performing traditional transfer learning model, EfficientNet, which achieved 97.18%. Furthermore, the proposed HCBLNet achieved competitive performance when compared with representative deepfake detection methods reported in the recent literature, demonstrating its effectiveness and robustness across multiple benchmark datasets. Our proposed model showed exceptional performance across different datasets, indicating its robustness and generalizability. For instance, on the FaceForensics++ dataset, our model achieved an accuracy of 95.40%, outperforming traditional models like ResNet152 and EfficientNet. Similarly, on the Celeb-DF dataset, our model achieved an accuracy of 95.97%, which is higher than the accuracy achieved by DenseNet201 and EfficientNet. These results underscore the model's capability to generalize well across various datasets, making it a reliable solution for deepfake detection in diverse scenarios.

When compared with existing deepfake detection systems, the proposed HCBLNet demonstrated competitive and consistent performance across the three benchmark datasets (DFDC, FaceForensics++, and Celeb-DF), highlighting its robustness and generalization capability. For example, the VGG16+LSTM model proposed by Boongasame et al. [34] achieved 96% accuracy on the Celeb-DF dataset, whereas

our model achieved a slightly lower accuracy of 95.97%, representing a marginal difference of only 0.28 percentage points. In contrast, Lopez Pellcier et al. [35] reported 95.1% accuracy on the Celeb-DF dataset using the DINOV2+xDNN model, while our model achieved a higher accuracy of 95.97%. Qadir et al. [36] reported 78.33% accuracy on DFDC and 95.23% on FaceForensics++, whereas our model achieved higher accuracies of 97.91% on DFDC and 95.40% on FaceForensics++. Omar et al. [37] achieved 97.81% accuracy on the Celeb-DF dataset using a convolution and self-attention network, which is higher than our model's 95.97%. Other notable comparisons include Ramadhani et al. [38], whose Video Vision Transformer (ViViT) achieved 88.03% on DFDC and 90.27% on FaceForensics++, and Liu et al. [39], whose DoubleStream+CViT model achieved 94.53% on DFDC and 95.10% on FaceForensics++. Overall, these comparisons indicate that HCBLNet provides strong and balanced performance across multiple benchmark datasets, achieving competitive or superior results in most cases while maintaining consistent generalization across diverse deepfake detection scenarios.

The training curves for our model on all three datasets demonstrated a consistent increase in accuracy and a decrease in loss over the epochs, indicating effective learning and generalization (Fig. 4). The 5-fold cross-validation results validated the robustness and reliability of our model, showing high average accuracies with low standard deviations across all folds, indicating stable performance regardless of the data split (Table 6). The confusion matrices and ROC curves provided additional insights into the model's performance, with minimal false positives and false negatives, and high precision and recall values (Figs. 5 and 6). The ROC curves showed excellent discriminative ability, with macro and micro AUC values exceeding 0.98, highlighting the model's effectiveness in accurately distinguishing between real and deepfake videos. Furthermore, we employed Grad-CAM and LIME X-AI to visualize the decision-making process of our proposed model. The Grad-CAM and LIME heatmaps (Fig. 7) revealed that the model consistently focused on semantically meaningful facial regions such as the eyes, mouth, and forehead when making predictions. This interpretability reinforces confidence in the model's predictions and confirms that it is learning relevant visual cues rather than spurious patterns.

The success of our model can be attributed to the integration of the multi-head attention mechanism with the CNN-BiLSTM architecture. The multi-head attention mechanism allows the model to focus on different parts of the input sequence simultaneously, capturing more nuanced features and improving the representation of the input data. This, combined with the feature extraction capabilities of the ResNet50 backbone and the sequential processing power of BiLSTM, results in a model that is highly effective at detecting deepfake videos. The algorithm of the proposed model is detailed in Algorithm 1.

Despite the promising results, there are several limitations to our study. First, the datasets used, while comprehensive, may not fully represent all possible real-world scenarios. For instance, our model tends to misclassify videos containing individuals with eyeglasses as fake, even when they are real. Similarly, inconsistencies in video quality or encoding artifacts can lead to false positives, where the model erroneously predicts a real video as fake. Second, the comparison with existing state-of-the-art methods is based on reported results in the literature, where differences in dataset splits, preprocessing pipelines, compression levels, and evaluation protocols may exist. This limits the direct one-to-one comparability of results across studies. Third, the model and hyperparameter selection process represents another limitation. Due to the experimental setup and the unavailability of a fully reproducible validation-based selection procedure, the test set may have been used, directly or indirectly, to inform model or hyperparameter selection. Therefore, the reported test-set performance should not be interpreted as a strictly unbiased held-out estimate of generalization performance.

Future work could address these limitations by evaluating the model on more diverse and challenging datasets, including videos with varying quality, compression levels, and different types of deepfake

manipulations. A dedicated validation set should also be used for model and hyperparameter selection, while an independent test set should be reserved exclusively for final evaluation to provide a more rigorous and unbiased assessment of generalization performance. Furthermore, exploring advanced architectures, such as transformers or graph neural networks, could potentially enhance the model's robustness and classification performance.

6 Conclusions

This study presents the development and evaluation of HCBLNet, a hybrid deep learning model integrating ResNet50, BiLSTM, and multi-head attention, for deepfake video detection. Our model achieved higher accuracy than several popular transfer learning-based models, including VGG16, ResNet, and EfficientNet, on three prominent datasets: DFDC, FaceForensics++, and Celeb-DF. Achieving accuracies of 97.91%, 95.40%, and 95.97%, respectively, the proposed model demonstrated solid and consistent performance on these benchmark datasets, comparable to or exceeding several existing methods reported in the literature. The integration of multi-head attention with CNN-BiLSTM architecture enabled our model to capture intricate features and deliver effective detection capabilities on the evaluated datasets. The training curves and cross-validation results support the model's reliability and consistency across the tested benchmarks. To enhance interpretability, we utilized explainable AI to visualize which regions of the face the model focused on during classification. The Grad-CAM and LIME heatmaps consistently highlighted key facial areas such as the eyes, mouth, and forehead, supporting the conclusion that the model relied on semantically meaningful features rather than irrelevant patterns. Despite these achievements, the model's performance is affected by specific limitations, such as misclassification of videos with eyeglasses and inconsistencies in video quality, as well as the inherent limitations of comparing results across studies with differing dataset splits, preprocessing pipelines, and evaluation protocols. Overall, HCBLNet should be regarded as a well-performing hybrid integration of established architectural design. Further evaluation on more diverse, real-world data is needed to substantiate claims of broader generalization. Future research should address these issues by incorporating more diverse datasets and exploring advanced techniques to further improve model accuracy, robustness, and generalization.

Acknowledgement: The authors would like to express their sincere gratitude to the Deanship of Graduate Studies and Scientific Research at Taif University and the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for their financial support and encouragement in completing this work.

Funding Statement: This work was supported by the Deanship of Graduate Studies and Scientific Research at Taif University. It was also supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) under grant number IMSIU-DDRSP2604.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Ashifur Rahman and Md. Rajaul Karim; methodology, Ashifur Rahman; software, Ashifur Rahman; validation, Ashifur Rahman, Md. Rajaul Karim and Amran Hossain; formal analysis, Ashifur Rahman; investigation, Ashifur Rahman; resources, Ashifur Rahman; data curation, Ashifur Rahman; writing—original draft preparation, Ashifur Rahman; writing—review and editing, Ashifur Rahman, Abdulmajeed M. Alenezi, Md. Shabiul Islam and Mohammad Tariqul Islam; visualization, Ashifur Rahman; supervision, Abdulmajeed M. Alenezi, Md. Shabiul Islam and Mohammad Tariqul Islam; project administration, Md. Rajaul Karim; funding acquisition, Abdulmajeed M. Alenezi, Md. Shabiul Islam and Mohammad Tariqul Islam. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in DFDC Dataset at https://www.kaggle.com/c/deepfake-detection-challenge/, FaceForensics++ at https://github.com/ondyari/FaceForensics, and Celeb-DF at https://github.com/yuezunli/celeb-deepfakeforensics.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, et al. Deep learning for deepfakes creation and detection: a survey. Comput Vis Image Underst. 2022;223:103525.
2. Abdulqudus A, Amodu O, Bukar U, Azlina R, Zakaria A, Queen SO, et al. A contemporary and comprehensive bibliometric exposition on deepfake research and trends. Comput Mater Contin. 2025;84(1):153–236. doi:10.32604/cmc.2025.061427.
3. George AS, George AH. Deepfakes: the evolution of hyper realistic media manipulation. Partn Univers Innov Res Publ. 2023;1(2):58–74.
4. Rahman A, Islam MM, Moon MJ, Tasnim T, Siddique N, Shahiduzzaman M, et al. A qualitative survey on deep learning based deep fake video creation and detection method. Aust J Eng Innov Technol. 2022;4(1):13–26. doi:10.34104/ajeit.022.013026.
5. Alrawahneh AAM, Abdullah SNAS, Abdullah SNHS, Kamarudin NH, Taylor SK. Video authentication detection using deep learning: a systematic literature review. Appl Intell. 2025;55(4):239. doi:10.1007/s10489-024-05997-8.
6. Westerlund M. The emergence of deepfake technology: a review. Technol Innov Manag Rev. 2019;9(11):39–52. doi:10.22215/timreview/1282.
7. Chesney B, Citron D. Deep fakes: a looming challenge for privacy, democracy, and national security. Calif L Rev. 2019;107:1753.
8. Ramadhani KN, Munir R. A comparative study of deepfake video detection method. In: Proceedings of the 2020 3rd International Conference on Information and Communications Technology (ICOIACT); 2020 Nov 24–25; Yogyakarta, Indonesia. p. 394–9.
9. Meskys E, Liaudanskas A, Kalpokiene J, Jurcys P. Regulating deep fakes: legal and ethical considerations. J Intellect Prop Law Pract. 2020;15(1):24–31. doi:10.1093/jiplp/jpz167.
10. Guo Y, Ke X, Ma J. A face replacement neural network for image and video. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing; 2019 Feb 22–24; Zhuhai, China. p. 163–7.
11. Arora T, Soni R. A review of techniques to detect the GAN-generated fake images. In: Generative adversarial networks for image-to-image translation. Amsterdam, The Netherlands: Elsevier; 2021. p. 125–59.
12. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. arXiv:1811.00656. 2018.
13. Jiang L, Li R, Wu W, Qian C, Loy CC. DeeperForensics-1.0: a large-scale dataset for real-world 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. 2889–98.
14. Zhao Z, Wang P, Lu W. Detecting deepfake video by learning two-level features with two-stream convolutional neural network. In: Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence; 2020 Apr 23–26; Tianjin, China. p. 291–7.
15. Passos LA, Jodas D, Costa KA, Souza Júnior LA, Rodrigues D, Del Ser J, et al. A review of deep learning-based approaches for deepfake content detection. Expert Syst. 2022;41(8):e13570. doi:10.22541/au.169735672.27713914/v1.
16. Alrawahneh AAM, Abdullah SNHS, Abuain T, Abdullah SNAS, Taylor SK, Suhaimi NHS. Decision-aid framework for face authentication detection using ResNext50 and BiLSTM to enhance media integrity. IEEE Access. 2025;13:89858–73. doi:10.1109/access.2025.3569792.
17. Rahman A, Siddique N, Moon MJ, Tasnim T, Islam M, Shahiduzzaman M, et al. Short and low resolution deepfake video detection using CNN. In: Proceedings of the 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC); 2022 Sep 16–18; Hyderabad, India. p. 259–64.
18. Gaashan SM, Kamarudin NH, Abdullah SNHS, Abdullah SNAS, Taylor SK. Development of enhanced video manipulation detection using hybrid CNN-LSTM method for object forgery. In: Proceedings of the 2025 International Conference on Advanced Computing Technologies (ICoACT); 2025 Mar 14–15; Sivalasi, India. p. 1–6.

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based Localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 618–26.
20. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397. 2020.
21. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1–11.
22. 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. 3207–16.
23. 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.
24. Ab Rahman B, Nazri MZA, Yaakub MR. A review of artificial intelligence techniques for combating fake news on social media. Int J Islam Appl Comput Sci Technol. 2024;12(4):29–42.
25. Bappy JH, Simons C, Nataraj L, Manjunath B, Roy-Chowdhury AK. Hybrid LSTM and encoder-decoder architecture for detection of image forgeries. IEEE Trans Image Process. 2019;28(7):3286–300. doi:10.1109/tip.2019.2895466.
26. Su L, Li C, Lai Y, Yang J. A fast forgery detection algorithm based on exponential-fourier moments for video region duplication. IEEE Trans Multimed. 2018;20(4):825–40. doi:10.1109/tmm.2017.2760098.
27. Jia S, Xu Z, Wang H, Feng C, Wang T. Coarse-to-fine copy-move forgery detection for video forensics. IEEE Access. 2018;6:25323–35. doi:10.1109/access.2018.2819624.
28. Su L, Li C. A novel passive forgery detection algorithm for video region duplication. Multidimens Syst Signal Process. 2018;29(3):1173–90. doi:10.1007/s11045-017-0496-6.
29. Zhao DN, Wang RK, Lu ZM. Inter-frame passive-blind forgery detection for video shot based on similarity analysis. Multimed Tools Appl. 2018;77(19):25389–408. doi:10.1007/s11042-018-5791-1.
30. Singh G, Singh K. Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation. Multimed Tools Appl. 2019;78(9):11527–62. doi:10.1007/s11042-018-6585-1.
31. Aloraini M, Sharifzadeh M, Agarwal C, Schonfeld D. Statistical sequential analysis for object-based video Forgery detection. Electron Imaging. 2019;31(5):1–7. doi:10.2352/issn.2470-1173.2019.5.mwsf-543.
32. Saddique M, Asghar K, Bajwa UI, Hussain M, Habib Z. Spatial video Forgery detection and localization using texture analysis of consecutive frames. Adv Electr Comput Eng. 2019;19(3):97–108. doi:10.4316/aece.2019.03012.
33. Huang CC, Zhang Y, Thing VL. Inter-frame video forgery detection based on multi-level subtraction approach for realistic video forensic applications. In: Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP); 2017 Aug 4–6; Singapore. p. 20–4.
34. Boongasame L, Boonpluk J, Soponmanee S, Muangprathub J, Thammarak K. Design and implement deepfake video detection using VGG-16 and long short-term memory. Appl Comput Intell Soft Comput. 2024;2024(1):8729440. doi:10.1155/2024/8729440.
35. Lopez Pellcier A, Li Y, Angelov P. PUDD: towards robust multi-modal prototype-based deepfake detection. arXiv:2406.15921. 2024.
36. Qadir A, Mahum R, El-Meligy MA, Ragab AE, AlSalman A, Awais M. An efficient deepfake video detection using robust deep learning. Heliyon. 2024;10(5):e25757. doi:10.1016/j.heliyon.2024.e25757.
37. Omar K, Sakr RH, Alrahmawy MF. An ensemble of CNNs with self-attention mechanism for DeepFake video detection. Neural Comput Appl. 2024;36(6):2749–65. doi:10.1007/s00521-023-09196-3.
38. Ramadhani KN, Munir R, Utama NP. Improving video vision transformer for Deepfake video detection using facial Landmark, Depthwise Separable Convolution and self attention. IEEE Access. 2024;12:8932–9. doi:10.1109/access.2024.3352890.
39. Liu W, She T, Liu J, Wang R, Yao D, Liang Z. Lips are lying: spotting the temporal inconsistency between audio and visual in lip-syncing DeepFakes. arXiv:2401.15668. 2024.

40. Wodajo D, Atnafu S. Deepfake video detection using convolutional vision transformer. arXiv:2102.11126. 2021.
41. Tan L, Wang Y, Wang J, Yang L, Chen X, Guo Y. Deepfake video detection via facial action dependencies estimation. Proc AAAI Conf Artif Intell. 2023;37(4):5276–84. doi:10.1609/aaai.v37i4.25658.
42. Ilyas H, Irtaza A, Javed A, Malik KM. Deepfakes examiner: an end-to-end deep learning model for Deepfakes videos detection. In: Proceedings of the 2022 16th International Conference on Open Source Systems and Technologies (ICOSST); 2022 Dec 14–15; Lahore, Pakistan. p. 1–6.
43. Wang T, Chow KP. Noise based Deepfake detection via multi-head relative-interaction. Proc AAAI Conf Artif Intell. 2023;37(12):14548–56. doi:10.1609/aaai.v37i12.26701.
44. Ahmad W, Ali I, Shahzad AS, Hashmi A, Ghaffar F. ResViT: a framework for Deepfake videos detection. Int J Electr Comput Eng Syst. 2022;13(9):807–13.
45. Cozzolino D, Rössler A, Thies J, Nießner M, Verdoliva L. ID-reveal: identity-aware DeepFake video detection. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 15088–97.
46. Wang T, Cheng H, Chow KP, Nie L. Deep convolutional pooling transformer for Deepfake detection. ACM Trans Multimed Comput Commun Appl. 2023;19(6):1–20. doi:10.1145/3588574.
47. Bhat NA, Giri KJ. DeepFake detection in images and videos: a survey on models, datasets, and evaluation metrics. Multimed Tools Appl. 2026;85(3):233.
48. Khan SA, Dang-Nguyen DT. Deepfake detection: analyzing model generalization across architectures, datasets, and pre-training paradigms. IEEE Access. 2024;12:1880–908. doi:10.1109/access.2023.3348450.
49. Bradski G. The opencv library. Dr Dobb's J Softw Tools Prof Program. 2000;25(11):120–3.
50. Geitgey A. Face recognition documentation. Release. 2019;1(3):3–37.
51. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. p. 1135–44.
52. Das AK, Naskar R. HTMDF-DD: hybrid triple modality based spatial-temporal features early fusion for deepfake detection. Comput Electr Eng. 2026;130:110860. doi:10.1016/j.compeleceng.2025.110860.
53. Zhang F, Shao C, Du K, Guo Y, Song P, Cao L, et al. Dual-granularity contrastive learning for DeepFake detection. IEEE Trans Inf Forensics Secur. 2026;21:1552–65. doi:10.1109/tifs.2026.3653569.
54. Song W, Guo S, Gao M, Li Q, Zhu X, Rida I. Deepfake detection via feature refinement and enhancement network. Image Vis Comput. 2025;162:105663. doi:10.1016/j.imavis.2025.105663.
55. Han R, Wang X, Bai N, Hou J, Zhang W, Li J. HSFF-Net: hierarchical spectral-feature fusion network for deepfake detection and localization. Neural Netw. 2025;192:107967. doi:10.1016/j.neunet.2025.107967.