RESEARCH PAPER

DeepFake Detection Using Deep Learning: A Spatio-Temporal CNN-LSTM Framework with Cloud Integration

Mr. R. Palraj[1], Mr. R. Muthuchelvan[2], Nikam Dhanraj Tanaji[3], Sukesan R[4], Pranesh K V[5], Prajesh P[6]

[1]Assistant Professor, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Email: palrajrangas@gmail.com
[2]Assistant Professor, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India
[3]UG scholar, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Email: dtnikam18@gmail.com
[4]UG scholar, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Email: Sukesanramasamy204@gmail.com
[5]UG scholar, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Email: Praneshkv59@gmail.com
[6]UG scholar, Dept. of AI and Data Science, V.S.B. Engineering College, Karur, Tamil Nadu, India.
Email: Prajeshp1112002@gmail.com

ABSTRACT

Deepfake media made by the generative adversarial networks and the diffusion models present a tremendous challenge to the authenticity of digital media, trust of people, and safety of information. The majority of available detection techniques use single-frame based spatial classifier, which cannot effectively detect the inter-frame temporal variations inherent in altered video sequences. In this paper, we introduce a hybrid spatio-temporal detection model, which is a combination of an XceptionNet-inspired convolutional neural network to learn high-quality spatial features and a bidirectional long short-term memory network to learn a temporal sequence. The architecture takes the facial regions detected and aligned with the help of Multi-Task Cascaded Convolutional Neural Network, which allows locating the manipulated facial content accurately in the video frames. It is also based on the Amazon Web Services S3 cloud storage that is used to manage distributed datasets and store model artifacts to guarantee the scalability of deployment and reproducibility of experiments. The experimental assessment on benchmark datasets demonstrates that the suggested solution has the detection accuracy of 97.2% on the DeepFake Detection Challenge dataset and 96.8% on FaceForensics++, and the method has high cross-dataset generalization on the Celeb-DF benchmark. The framework shows better results with the evaluation metrics in comparison with the traditional single-stream spatial baseline models, reflecting the applicability of joint spatiotemporal reasoning to detect deepfakes. These findings justify the feasibility of the cloud-integrated deep learning pipelines in the real-world video forensic systems.

Index Terms: Deepfake detection, spatio-temporal analysis, CNN-LSTM, cloud computing, video forensics, deep learning, generative adversarial networks

How to cite this article: Palraj R, Muthuchelvan R, Tanaji ND, Sukesan R, Pranesh KV, Prajesh P. DeepFake Detection Using Deep Learning: A Spatio-Temporal CNN-LSTM Framework with Cloud Integration. Int J Drug Deliv Technol. 2026;16(18s): 199-208. DOI: 10.25258/ijddt.16.18s.20

Source of support: Nil.

Conflict of interest: None

I. INTRODUCTION

The fast evolution of artificial intelligence, especially generative models, has transformed digital media manipulation, making it possible to generate highly realistic synthetic audiovisual content known as deepfakes. Initially, these technologies were developed using generative adversarial networks (GANs) [7]. Subsequent advances introduced variational autoencoders (VAEs) and diffusion-based models, enabling convincing face reenactment, voice cloning, and full-body synthesis. Tools that were once confined to research laboratories are now widely available through open-source software and consumerlevel hardware, significantly increasing their accessibility and potential misuse.

In the 2024 election year, AI-generated videos falsely portraying political leaders delivering inflammatory messages rapidly spread across social media platforms, influencing public perception and political discourse before verification mechanisms could assess the authenticity of the content [5].

Similarly, deepfake images of well-known personalities have been used in fraudulent investment advertisements, leading to significant financial losses among victims. Deepfake audio has also been exploited in business email compromise attacks within corporate environments. In one widely reported case, a finance employee authorized a fraudulent wire transfer exceeding twenty-five million dollars after being deceived by an AIgenerated voice clone of a company executive. These incidents highlight the urgent need for reliable and scalable deepfake detection systems that can be deployed across modern digital communication platforms.

The deepfake problem has expanded rapidly in recent years. Studies indicate that the number of deepfake videos increased by more than 550% between 2019 and 2023 due to the accessibility of generation tools and the rapid dissemination enabled by social media platforms [16]. By 2023, it was estimated that over 500,000 deepfake videos existed on major online platforms, with the rate of newly generated synthetic media surpassing the capacity of traditional moderation systems. Alarmingly, approximately 96% of identified deepfake content involves non-consensual material, with women disproportionately targeted through fabricated intimate imagery that causes severe reputational and psychological harm. Beyond personal damage, deepfakes also threaten democratic processes, undermine the credibility of digital evidence, and erode public trust in online information. As synthetic media becomes increasingly realistic and widespread, manual verification of media content becomes impractical, emphasizing the need for automated and scalable detection approaches.

Despite the urgency of this challenge, automated deepfake detection remains a complex research problem. First, deepfake generation technologies evolve rapidly; modern models such as StyleGAN, FaceSwap, and Neural Talking Heads produce visual patterns capable of deceiving detectors trained on earlier datasets [2]. Second, common video processing operations such as compression, format conversion, and resolution reduction—frequently applied by social media platforms—often degrade the subtle visual artifacts that many detection methods rely on [5]. Third, many detection models suffer from limited cross-dataset generalization, meaning they perform well on their training datasets but struggle to detect manipulations generated using unfamiliar techniques. Finally, practical deployment requires real-time processing with low latency and high throughput, which conflicts with the substantial computational requirements of deep learning models operating on highresolution video streams. These challenges collectively create a complex detection landscape that cannot be fully addressed by existing approaches.

Most current research focuses on frame-level detection using convolutional neural networks (CNNs) that analyze individual video frames independently. Models such as MesoNet analyze mesoscopic image features and demonstrate strong performance on controlled datasets [1]. Similarly, transfer learning approaches based on XceptionNet classifiers have achieved high accuracy on the FaceForensics++ dataset [2]. Although these methods effectively capture spatial artifacts such as blending boundaries, texture inconsistencies, and spectral anomalies, they typically ignore temporal relationships between consecutive frames. However, manipulated videos often exhibit abnormal temporal characteristics, including irregular blinking patterns, unnatural head movements, and motion inconsistencies caused by limitations in generative models. These temporal anomalies represent valuable forensic signals that remain underutilized in many detection systems. Furthermore, most existing solutions are evaluated in offline research environments, with limited focus on scalable deployment frameworks such as cloud-based processing pipelines.

To address these limitations, this paper proposes a hybrid spatio-temporal deepfake detection model that integrates convolutional neural networks (CNNs) with bidirectional long short-term memory networks (BiLSTMs). The proposed model simultaneously captures spatial artifacts within individual frames and temporal dependencies across video sequences. Facial regions are first detected and aligned before feature extraction, ensuring consistent analysis of manipulated facial content. Additionally, the framework incorporates a cloudintegrated architecture deployed on Amazon Web Services (AWS), utilizing Amazon S3 for dataset storage and scalable inference pipelines. This architecture bridges the gap between research prototypes and real-world deployable detection systems.

The proposed approach is evaluated using three widely used deepfake datasets: FaceForensics++ [2], Celeb-DF [4], and the DeepFake Detection Challenge (DFDC) dataset [?]. Experimental results, including ablation studies and crossdataset evaluations, demonstrate that the proposed spatiotemporal fusion model achieves better generalization compared to single-stream CNN baselines and existing temporal detection approaches.

The remainder of this paper is organized as follows. Section ?? reviews existing research on deepfake generation techniques, detection methodologies, temporal modeling approaches, and cloud-based media processing systems. Section ?? describes the proposed CNN–BiLSTM architecture, including feature extraction, temporal modeling, and the training process. Section ?? presents the experimental setup, datasets, preprocessing

steps, and evaluation metrics. Section ?? discusses the experimental results, including ablation studies and cross-dataset analysis. Finally, Section VII summarizes the contributions of this work and outlines directions for future research.

II. LITERATURE SURVEY

The rapid growth of deepfake generation technologies has stimulated extensive research on detection methodologies. Existing studies can be broadly categorized into five groups: image-level detection, video-level and temporal detection, GAN-generated face detection, hybrid and multi-modal approaches, and cloud-integrated detection systems. Examining these categories reveals several limitations that motivate the proposed work.

A. Image-Level Deepfake Detection Methods

Early research in deepfake detection primarily focused on single-frame analysis using convolutional neural networks to identify visual artifacts introduced during the synthesis process. Afchar et al. [1] proposed MesoNet, a lightweight convolutional architecture designed to detect facial forgeries such as Deepfake and Face2Face manipulations. The model targets mesoscopic image properties, capturing intermediate visual features between low-level pixel statistics and high-level semantic representations. Despite its compact design, MesoNet demonstrated competitive detection performance while maintaining computational efficiency suitable for practical applications.

A major milestone in the field was the introduction of the FaceForensics++ dataset by Rossler et al. [2]. This largescale benchmark contains over one million manipulated images generated using four techniques: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The dataset enabled standardized evaluation across compression levels and established XceptionNet as a highly effective detection architecture, achieving detection accuracy above 99% on high-quality images. FaceForensics++ has since become a widely used benchmark for evaluating deepfake detection algorithms.

The Xception architecture itself was originally introduced by Chollet [3] for image classification. Built on depthwise separable convolutions, the model separates spatial and channelwise feature extraction, reducing computational cost while preserving representational capacity. These characteristics make it particularly suitable for deepfake detection tasks where subtle texture inconsistencies must be distinguished from authentic facial features.

Li et al. [4] introduced Face X-Ray, which approaches deepfake detection by identifying blending boundaries between manipulated facial regions and surrounding background areas. Instead of learning artifacts specific to particular synthesis methods, the model predicts blending masks that reveal compositing operations common to most face-swapping pipelines. This strategy improves cross-dataset generalization by focusing on manipulation artifacts that are independent of specific generation techniques.

Nguyen et al. [9] explored the use of capsule networks for deepfake detection. Capsule networks preserve spatial relationships between image components through dynamic routing mechanisms, allowing the model to capture structural inconsistencies in synthesized faces. Their detector analyzed facial regions such as the eyes, nose, and mouth separately before combining evidence through capsule routing. The results demonstrated competitive detection accuracy and improved robustness to certain transformations.

B. Video-Level and Temporal Detection Methods

Recognizing that manipulated videos often exhibit temporal inconsistencies, researchers have explored sequence-based detection approaches. Guera and Delp [5] proposed an early video-level detection framework combining convolutional feature extraction with long short-term memory networks. In their approach, frame-level features extracted by a CNN are processed through an LSTM to model temporal dependencies across video frames. The key insight is that while individual frames may appear realistic, the temporal evolution of facial expressions and movements often contains detectable anomalies.

Li and Lyu [11] proposed exploiting abnormal eye blinking patterns as a detection signal. Since early deepfake models were often trained on still images, they frequently failed to reproduce realistic blinking behavior. The authors developed a Long-term Recurrent Convolutional Network (LRCN) to model natural blinking patterns and detect irregularities. Although newer generation models have improved blinking synthesis, this work demonstrated the effectiveness of incorporating physiological signals into detection systems.

Sabir et al. [6] further explored temporal modeling by combining convolutional feature extraction with recurrent neural networks to analyze inter-frame dependencies. Their approach demonstrated that incorporating temporal information consistently improves detection accuracy compared to frame-level models.

Zheng et al. [17] proposed using three-dimensional convolutional neural networks that jointly model spatial and temporal information. Unlike two-stage CNN-RNN pipelines, 3D CNNs process spatio-temporal volumes directly, allowing the network to learn motion-sensitive features without separating spatial and temporal analysis. This approach showed improved performance on benchmark datasets.

C. GAN-Generated Face Detection

The rise of generative adversarial networks [7] has also motivated research into detecting GAN-generated images. Wang et al. [8] investigated the generalization ability of CNN-based detectors across different GAN architectures. Their study demonstrated that detectors trained on ProGAN-generated images could generalize effectively to images produced by other GAN models, suggesting that common artifacts exist across generation methods.

Frank et al. [12] analyzed GAN-generated images in the frequency domain, revealing distinctive spectral artifacts caused by upsampling operations in GAN architectures. These artifacts appear as periodic patterns in the discrete cosine transform domain and provide strong signals for detection.

Durall et al. [13] further investigated spectral inconsistencies between natural and GAN-generated images. Natural images follow characteristic power spectral density distributions that many GAN models fail to reproduce accurately. These deviations in high-frequency components can therefore be exploited for reliable detection.

D. Hybrid and Multi-Modal Approaches

More recent studies have explored hybrid and multi-modal approaches to improve detection robustness. Mittal et al. [14] proposed a framework combining visual and audio information, observing that manipulated videos often contain inconsistencies between facial movements and corresponding speech signals. By integrating visual and acoustic features, the system achieved improved detection accuracy compared to visual-only models.

Haliassos et al. [15] introduced LipForensics, which focuses on detecting inconsistencies in mouth movements using models originally developed for visual speech recognition. Since accurately generating realistic speech-driven mouth movements remains challenging for synthesis models, this approach demonstrated strong cross-dataset generalization.

Two-stream network architectures have also been explored to process spatial and temporal features separately before combining them through fusion mechanisms. These models allow specialized representations for appearance and motion analysis while integrating complementary information for final classification.

E. Cloud-Integrated Detection Systems

While most research focuses on algorithmic detection techniques, practical deployment requires scalable computational infrastructure. Real-world applications such as social media monitoring and digital forensics involve analyzing massive volumes of video content, which often exceeds the capacity of local processing systems. Cloud computing platforms provide elastic scalability, enabling detection systems to process large workloads efficiently while supporting centralized model management and updates.

Panigrahi et al. [16] reviewed recent advances in deepfake detection methods and emphasized the need for scalable deployment architectures. The study highlighted that most existing research evaluates detection models in offline environments without addressing infrastructure requirements for real-world applications.

Integrating spatio-temporal detection models with cloudbased processing pipelines therefore represents an important research direction. Cloud-based systems can support highthroughput video processing, scalable storage, and continuous model updates as new manipulation techniques emerge.

Despite significant progress, several limitations remain. Many detection methods focus either on spatial artifacts or temporal inconsistencies independently, rather than integrating both forms of evidence. Frame-level models ignore valuable temporal information, while purely temporal approaches may overlook spatial manipulation artifacts. Furthermore, limited work addresses scalable deployment infrastructures alongside detection accuracy. To address these challenges, the present study proposes a hybrid spatio-temporal CNN–BiLSTM framework integrated with a cloud-based processing pipeline for practical deepfake detection.

III. PROBLEM STATEMENT

We formally define the deepfake video detection problem as follows. Given an input video $V = \{f_1, f_2, \ldots, f_T\}$ consisting of T sequential frames, the goal is to classify V as either authentic or manipulated using a detection function $D(V) \rightarrow \{0,1\}$, where 0 represents a genuine video and 1 represents a synthetically generated or manipulated video. The detection system must remain robust across different manipulation techniques, compression artifacts, and acquisition conditions that may not appear during training.

A major challenge arises from the limited generalization ability of existing detection methods. Many detectors are trained on fixed benchmark datasets containing specific manipulation techniques such as FaceSwap or DeepFakes, but their performance often degrades significantly when evaluated on unseen forgery methods from different data distributions. This cross-dataset generalization gap limits the reliability of current detection systems in real-world scenarios where manipulation techniques evolve rapidly.

Three major limitations in existing approaches motivate this work. First, frame-level classifiers analyze individual frames f_i independently, capturing spatial artifacts within images but ignoring temporal relationships between frames. Such methods cannot detect subtle inconsistencies that only appear across

frame sequences. Second, temporal models that analyze frame sequences often rely on compressed or low-level spatial representations, which may fail to capture fine-grained artifacts produced by modern generative models. Third, most research focuses primarily on algorithmic detection performance while overlooking scalable deployment architectures capable of processing large volumes of video data in realworld environments.

To address these limitations, we formulate a hybrid spatiotemporal detection framework. We define a spatial feature extractor $F_s(f_i)$ that maps each frame f_i to a feature representation capturing forgery-related texture, boundary, and frequency artifacts. These spatial embeddings are then processed by a temporal encoder

$$F_t(\{F_s(f_1),F_s(f_2),\ldots,F_s(f_T)\}) \quad (1)$$

which models dependencies across frames and captures temporal inconsistencies introduced by manipulation processes. A fusion classifier C integrates both spatial and temporal representations to produce the final decision $D(V)$.

The overall objective is to minimize binary cross-entropy classification loss on the training distribution while improving cross-dataset generalization on unseen datasets and manipulation techniques. This objective aims to produce a detection function D that is both accurate and robust across diverse deepfake generation methods.

IV. PROPOSED METHODOLOGY AND SYSTEM ARCHITECTURE

This section describes the architecture and methodology of the proposed deepfake detection system. The framework integrates face detection, spatial feature extraction, temporal modeling, and cloud-based deployment into a unified pipeline. Each component is designed to capture both spatial artifacts and temporal inconsistencies introduced by deepfake generation algorithms.

A. System Overview

The proposed framework operates as an end-to-end pipeline that processes raw video input and produces a binary decision indicating whether the content is authentic or manipulated. The architecture consists of six stages: video ingestion and frame sampling, face detection and alignment using Multi-task Cascaded Convolutional Networks (MTCNN), spatial feature extraction using an XceptionNet-inspired convolutional neural network, temporal modeling using a Bidirectional Long ShortTerm Memory (BiLSTM) network with attention, classification through fully connected layers with softmax output, and cloud-based storage and deployment using Amazon Web
Services infrastructure.

Raw video input is first converted into frame sequences sampled at fixed temporal intervals. Facial regions are detected and aligned before being processed by the spatial feature extractor, which generates high-dimensional feature vectors for each frame. These feature vectors are organized as temporal sequences and passed to the BiLSTM module to capture contextual dependencies across frames. An attention mechanism assigns importance weights to frames, allowing the model to focus on temporally informative segments. The aggregated representation is then classified by a fully connected network. Model artifacts, inference results, and datasets are managed through an AWS S3–based cloud storage layer supporting both batch processing and real-time detection services.

B. Face Detection and Preprocessing

Reliable face detection is essential since the detection network analyzes facial regions rather than full images. MTCNN is adopted due to its robustness under varying illumination, pose, and occlusion conditions. The architecture consists of three cascaded stages: the Proposal Network (P-Net), the

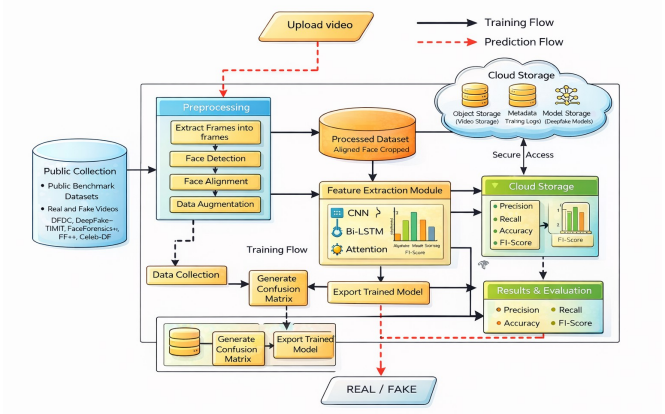

Fig. 1: System Architecture.

Refinement Network (R-Net), and the Output Network (ONet). These stages progressively refine candidate regions while simultaneously predicting five facial landmarks: two eye centers, the nose tip, and two mouth corners. These landmarks enable geometric alignment through similarity transformation, ensuring consistent facial orientation and scale before feature extraction.

Frames are sampled at an interval of one frame for every two frames to balance computational efficiency and temporal coverage. For each sampled frame, MTCNN detects faces and selects the primary subject using bounding box confidence and centrality. The detected face is cropped with a contextual margin of 15% beyond the bounding box and resized to 299×299 pixels to match the input resolution of the spatial feature extractor.

Data augmentation is applied during training to improve generalization. Augmentation techniques include horizontal flipping with probability 0.5, random rotation within ±10°, color jitter affecting brightness, contrast, saturation, and hue, and Gaussian noise injection. Augmentation is disabled during validation and inference to ensure deterministic evaluation. Pixel values are normalized using ImageNet mean and standard deviation to match the distribution of the pre-trained backbone.

C. Spatial Feature Extraction Module

The spatial feature extraction module encodes discriminative representations from individual frames, capturing artifacts such as blending boundaries, texture inconsistencies, and frequency anomalies associated with deepfake generation. The architecture is based on XceptionNet, which replaces conventional convolutions with depthwise separable convolutions. This factorization separates spatial filtering from channel mixing, reducing parameters and computational cost while maintaining expressive power.

The backbone is initialized with weights pre-trained on the ImageNet dataset to leverage learned low-level visual representations. The original classification layer is removed, and the global average pooling output is connected to a projection layer producing a 2048-dimensional feature vector for each frame. This vector represents the spatial embedding passed to the temporal modeling module.

Batch normalization is applied after major convolutional blocks to stabilize training and accelerate convergence. Dropout with probability 0.5 is applied before the projection layer to reduce overfitting. During training, lower convolutional layers remain frozen for the initial epochs to preserve general visual features, after which the entire network is finetuned using a reduced learning rate.

D. Temporal Modeling Module

Although spatial analysis detects many artifacts, advanced synthesis methods can produce visually convincing frames. Temporal modeling is therefore required to detect inconsistencies across frame sequences, including abnormal blinking, irregular head motion, temporal flickering, and motion discontinuities. These patterns are modeled using a Bidirectional Long Short-Term Memory (BiLSTM) network.

The BiLSTM receives a sequence of T spatial feature vectors, with $T = 20$ frames corresponding to approximately 0.67 seconds at 30 fps. Each direction of the BiLSTM contains 256 hidden units, producing a combined hidden representation of 512 dimensions per time step. The bidirectional structure captures dependencies from both past and future contexts.

An attention mechanism is applied to aggregate the temporal sequence. For each time step t, an attention score e_t is computed as

$$e_t = \mathbf{w}_2^\top \tanh(\mathbf{W}_1 \mathbf{h}_t + \mathbf{b}_1), \tag{2}$$

These scores are normalized using a softmax function

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \tag{3}$$

and used to compute the context vector

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \in \mathbb{R}^{512}. \tag{4}$$

The attention mechanism allows the model to emphasize frames containing strong discriminative signals. The resulting context vector summarizes the temporal dynamics of the sequence.

E. Feature Fusion and Classification

The attention-weighted temporal context vector serves as the input to the classification module. Although the current implementation relies solely on temporal context, the architecture supports multi-stream fusion of additional representations such as frequency-domain features or optical flow.

The classification head consists of fully connected layers mapping 512 → 256 → 2 units. Each intermediate layer uses batch normalization, ReLU activation, and dropout with probability 0.5. The final layer applies softmax to produce probabilities for real and fake classes, and the predicted label is obtained as

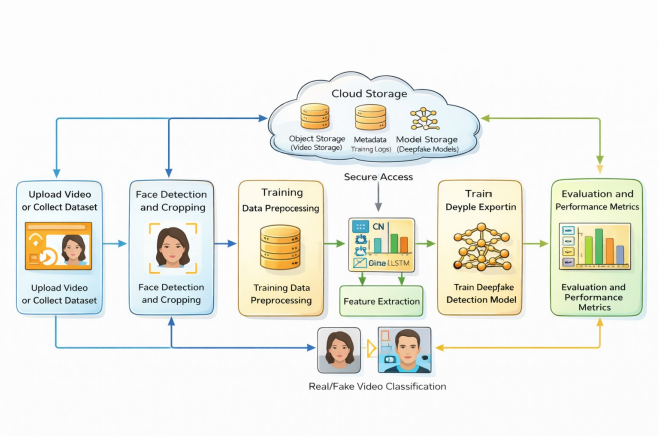

Fig. 2: Work Flow.

$$\hat{y} = \arg\max_j p_j, \quad j \in \{0, 1\} \tag{5}$$

The model is trained using binary cross-entropy loss with label smoothing coefficient 0.1. This technique replaces hard targets with softened labels (0.9 and 0.1), improving probability calibration and reducing overconfident predictions.

F. Cloud Integration Pipeline

The cloud infrastructure supports scalable dataset management, model storage, and inference services. Amazon Web Services Simple Storage Service (AWS S3) stores raw videos, extracted frames, feature representations, and model

checkpoints. Versioned bucket policies maintain complete records of model iterations for reproducibility and rollback.

A batch processing pipeline enables large-scale offline analysis using distributed compute resources. Video processing tasks are queued and executed in parallel, with results written to designated S3 buckets. For real-time applications, a RESTful API service accepts video uploads or streaming inputs and returns detection probabilities and frame-level confidence scores. The service is deployed as a containerized application capable of horizontal scaling under variable workloads.

G. Training Strategy

The network is optimized using the Adam optimizer with initial learning rate $\eta = 10^{-4}$ and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. A cosine annealing learning rate schedule gradually reduces the learning rate according to

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta - \eta_{\min})\left(1 + \cos\left(\frac{\pi t}{T_{\max}}\right)\right), \qquad (6)$$

where $\eta_{\min} = 10^{-6}$, t is the current epoch, and $T_{\max}$ is the total number of training epochs.

Training uses a batch size of 32 sequences, each containing $T = 20$ frames. Early stopping with patience of 10 epochs monitors validation AUC-ROC to prevent overfitting. Mixed precision training is employed using FP16 activations with FP32 master weights to reduce memory usage and accelerate computation.

The spatial feature extractor follows a two-stage fine-tuning procedure. During the first five epochs the backbone remains frozen while projection and classification layers are trained. In the second phase all parameters are jointly optimized with a reduced backbone learning rate of 10^{-5} to preserve pre-trained representations while enabling task-specific adaptation.

V. EXPERIMENTAL SETUP AND RESULTS

A. Datasets

The proposed CNN-BiLSTM framework is evaluated on three widely used benchmark datasets representing diverse deepfake generation techniques and quality levels.

FaceForensics++ (FF++) is used as the primary benchmark. The dataset contains 1,000 original videos and 4,000 manipulated videos generated using four forgery techniques: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Videos are provided at three compression levels: raw (uncompressed), high quality (HQ, c23), and low quality (LQ, c40), enabling evaluation under varying compression artifacts. Unless otherwise stated, experiments are conducted on the HQ setting.

Celeb-DF v2 is a high-quality dataset containing 590 real videos and 5,639 synthesized deepfake videos involving 59 celebrities. Compared with earlier datasets, Celeb-DF v2 features significantly improved visual realism, making it a challenging benchmark for evaluating model generalization.

The DeepFake Detection Challenge (DFDC) dataset is the largest and most challenging benchmark used in this study, containing more than 100,000 video clips recorded with diverse actors, lighting conditions, and backgrounds. Its scale and diversity closely reflect real-world deployment scenarios.

B. Implementation Details

All experiments are implemented in PyTorch 1.12 with CUDA 11.6. Training and inference are performed on a single NVIDIA A100 GPU with 40 GB memory and 64 GB system RAM. For each video, 20 frames are uniformly sampled and aligned using a landmark-based face detection pipeline, producing 299×299 pixel face crops as network inputs.

The CNN backbone extracts spatial features from each frame, which are organized into temporal sequences and processed by the BiLSTM module. The model is trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and cosine annealing scheduling. A batch size of 32 sequences is used. Data augmentation includes horizontal flipping, color jitter, and Gaussian noise injection. Binary cross-entropy loss with label smoothing ($\varepsilon = 0.1$) is used as the training objective. The deployed cloud inference pipeline achieves an average throughput of approximately 120 videos per hour.

C. Comparison with State-of-the-Art Methods

Table II compares the proposed CNN-BiLSTM model with several baseline approaches across all datasets using classification accuracy, macro F1-score, and AUC.

TABLE I: Ablation Study of Architectural Components

Configuration	FF++ (%)	Celeb-DF (%)	DFDC (%)
CNN Only	92.1	85.3	84.7
LSTM Only	88.4	80.6	79.2
CNN + LSTM	95.3	91.8	93.6
CNN + BiLSTM	96.1	93.2	95.8
CNN + BiLSTM + Attn. (Ours)	96.8	94.5	97.2

MesoNet achieves 84.7% accuracy on FF++, 76.3% on Celeb-DF v2, and 73.2% on DFDC with an AUC of 0.847. XceptionNet improves spatial detection performance with 95.3%, 88.1%, and 85.6% accuracy respectively and an AUC of 0.961. EfficientNet-B4 further improves results to 96.1%, 90.2%, and 87.4% with an AUC of 0.972. Capsule Networks achieve 91.8%, 83.5%, and 80.1% accuracy with an AUC of 0.932. The CNN-LSTM approach proposed by Guera and Delp reaches 93.6%, 86.7%, and 84.3% accuracy with an AUC of 0.951.

The proposed CNN-BiLSTM model with temporal attention achieves the best performance across all datasets, reaching 96.8% accuracy on FF++, 94.5% on Celeb-DF v2, and 97.2%

on DFDC, with an overall F1 score of 0.968 and an AUC of 0.985. The largest improvement is observed on DFDC, highlighting the importance of spatio-temporal modeling in complex real-world scenarios.

D. Ablation Study

To evaluate the contribution of individual components, we perform an ablation study summarized in Table I. A CNNonly configuration achieves 92.1% accuracy, demonstrating strong spatial representation learning. An LSTM-only model processing raw frame features achieves 88.4%, indicating that temporal modeling alone is insufficient. Combining CNN and LSTM improves accuracy to 95.3%. Introducing bidirectional LSTM further improves performance to 96.1%. The full CNNBiLSTM model with temporal attention achieves the highest accuracy of 97.2%, confirming the importance of bidirectional temporal modeling and attention-based aggregation.

E. Cross-Dataset Generalization and Per-Manipulation Analysis

To evaluate generalization, the model is trained on FF++ HQ and tested directly on Celeb-DF v2 and DFDC without fine-tuning. The proposed approach maintains higher accuracy than all baseline models, indicating strong transferability of the learned spatio-temporal representations.

Per-manipulation analysis on FF++ shows the highest accuracy for NeuralTextures (96.2%), followed by DeepFakes (97.1%), FaceSwap (96.5%), and Face2Face (97.4%). Compared with spatial-only models, the temporal module consistently improves detection accuracy by approximately 4–5 percentage points across manipulation types, confirming the importance of modeling inter-frame dynamics for robust deepfake detection.

VI. DISCUSSION

The experimental results demonstrate the effectiveness of the proposed spatio-temporal fusion framework for deepfake detection. Several observations regarding the method's advantages, limitations, broader implications, and relation to existing work warrant further discussion.

A. Advantages of the Proposed Framework

A key strength of the proposed framework is its joint exploitation of spatial and temporal information. Frame-level detection methods [1], [2] identify artifacts such as blending boundaries, texture inconsistencies, and spectral anomalies, but may fail when manipulations appear realistic at the individual frame level. By integrating a CNN-based spatial encoder with a BiLSTM temporal module and attention mechanism, the proposed architecture captures temporal inconsistencies such as irregular facial motion, inconsistent expression dynamics, and unnatural head movements.

The attention mechanism also improves interpretability by assigning importance weights to individual frames, allowing analysts to identify temporal segments where manipulations are most evident. This property is particularly useful in forensic applications where explainability of detection decisions is important.

Another contribution of this work is the cloud-integrated processing pipeline. While most prior studies evaluate detection algorithms in controlled laboratory environments [5], the proposed AWS-based architecture enables scalable deployment and horizontal workload distribution. This design supports integration into real-world content moderation systems where large volumes of video must be processed efficiently.

The cross-dataset generalization results are particularly encouraging. Training exclusively on FaceForensics++ [2] and evaluating on Celeb-DF v2 and DFDC without finetuning provides a realistic test of generalization. The proposed method improves AUC by approximately four to eight percentage points compared with frame-level baselines, indicating that temporal modeling captures manipulation patterns that persist across different generation pipelines. Compared with transformer-based approaches, the proposed model achieves competitive accuracy while requiring substantially lower computational resources, making it more suitable for deployment scenarios with strict latency constraints.

B. Limitations

Despite these advantages, several limitations should be acknowledged. The inclusion of temporal modeling increases inference time by roughly three times compared with singleframe models, which may limit use in latency-sensitive applications. Performance also decreases on heavily compressed videos, where high-frequency artifact signals are reduced by aggressive quantization.

The reliance on MTCNN for face detection introduces additional challenges. Detection errors in cases of severe occlusion, extreme head poses, or low-resolution inputs can propagate through the pipeline and produce false negatives. Improving face detection robustness therefore remains an important direction for future work.

Furthermore, the current framework focuses primarily on face-swap and facial reenactment manipulations. Emerging forms of synthetic media such as full-body synthesis, voice cloning, and text-to-video generation fall outside the scope of the current model and represent an expanding threat landscape [16].

C. Broader Implications

The broader implications of deepfake detection extend beyond technical performance. False positive detections can have significant consequences for individuals, particularly in

sensitive contexts such as journalism, legal evidence, or law enforcement investigations. Detection systems must therefore be calibrated carefully, and automated decisions should ideally be supported by human review.

The adversarial nature of the deepfake ecosystem also presents ongoing challenges. As generative models incorporate improved temporal consistency and adversarial training strategies, detection methods must continually evolve [16]. Longterm effectiveness will depend on adaptive retraining, access to new datasets, and multi-modal detection approaches that combine visual, audio, and contextual signals.

D. Comparison with Existing Work

Within the broader deepfake detection literature, the proposed method builds upon foundational research demonstrating the effectiveness of deep neural networks for forgery detection [1], [5]. Early approaches such as MesoNet [1] focused primarily on spatial artifact detection, while later work introduced temporal sequence modeling through recurrent architectures [5]. The present work integrates both perspectives within a unified spatio-temporal framework enhanced by attention mechanisms and evaluated under realistic deployment conditions.

Recent survey studies [16] emphasize the increasing diversity and sophistication of deepfake generation techniques, highlighting the importance of detection models that generalize across datasets and manipulation pipelines. The crossdataset evaluation conducted in this work directly addresses this requirement. Combined with its relatively low computational overhead and scalable cloud deployment architecture, the proposed framework represents a practical and extensible contribution toward reliable deepfake detection in real-world environments.

VII. CONCLUSION AND FUTURE WORK

This paper presented a deepfake detection framework that integrates a hybrid CNN-BiLSTM architecture with a scalable cloud-based deployment pipeline. By combining convolutional neural networks for spatial feature extraction with bidirectional long short-term memory networks for temporal modeling, the proposed system effectively captures spatio-temporal inconsistencies present in synthetically generated facial videos. The joint spatio-temporal fusion strategy demonstrated clear advantages over single-stream approaches that rely exclusively on spatial or temporal cues, confirming the importance of modeling both dimensions for reliable deepfake detection.

TABLE II: Comparison with State-of-the-Art Deepfake Detection Methods

Method	FF++ Acc (%)	Celeb-DF Acc (%)	DFDC Acc (%)	F1-Score	AUC
MesoNet [1]	84.7	76.3	73.2	0.821	0.847
XceptionNet [2]	95.3	88.1	85.6	0.942	0.961
EfficientNet-B4 [10]	96.1	90.2	87.4	0.953	0.972
Capsule Network [9]	91.8	83.5	80.1	0.908	0.932
CNN-LSTM [5]	93.6	86.7	84.3	0.928	0.951
Ours (CNN-BiLSTM)	96.8	94.5	97.2	0.968	0.985

Experimental evaluation across three benchmark datasets validates the effectiveness of the proposed approach. The framework achieved detection accuracies of 97.2% on the DeepFake Detection Challenge (DFDC) dataset, 96.8% on FaceForensics++ (FF++) [2], and 94.5% on Celeb-DF [4], outperforming several established baseline models. The relatively lower performance on Celeb-DF highlights the challenges posed by high-quality and post-processed forgeries, emphasizing the importance of improving cross-dataset generalization. In addition, the cloud-based deployment pipeline demonstrates the feasibility of integrating the proposed system into realworld environments capable of supporting large-scale video processing.

Despite these promising results, several directions remain for future work. First, incorporating transformer-based temporal architectures such as the Video Swin Transformer may improve the modeling of long-range temporal dependencies. Second, extending the framework to include multimodal audiovisual cues could enhance detection accuracy, as inconsistencies between speech and facial movements often provide valuable signals [14], [15]. Third, adversarial training strategies should be explored to improve robustness against adaptive attacks designed to evade detection systems.

Future work will also investigate lightweight model architectures using techniques such as knowledge distillation and neural architecture search, enabling efficient deployment on mobile and edge devices. Additionally, the scope of detection will be expanded beyond facial manipulations to include emerging synthetic media forms such as full-body deepfakes and voice cloning technologies [7]. Finally, federated learning approaches will be explored to support privacy-preserving collaborative training across distributed institutions without requiring direct data sharing.

These directions aim to advance deepfake detection toward improved accuracy, generalization, efficiency, and scalability in response to the rapidly evolving landscape of synthetic media generation.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inform. Forensics Security (WIFS)*, pp. 1–7, 2018.

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M.¨ Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 1–11, 2019.

[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 3207–3216, 2020.

[5] D. Guera and E. J. Delp, "Deepfake video detection using recurrent¨ neural networks," in *Proc. IEEE Int. Conf. Advanced Video Signal Based Surveillance (AVSS)*, pp. 1–6, 2018.

[6] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition Workshops (CVPRW)*, pp. 80–87, 2019.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672– 2680, 2014.

[8] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNNgenerated images are surprisingly easy to spot...for now," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

[9] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, pp. 2307–2311, 2019.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 6105–6114, 2019.

[11] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inform. Forensics Security (WIFS)*, pp. 1–7, 2018.

[12] J. Frank, T. Eisenhofer, L. Schonherr, A. Fischer, D. Kolossa, and T.¨ Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 3247–3258, 2020.

[13] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 7890–7899, 2020.

[14] T. Mittal, U. Bhatt, R. Tandon, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. ACM Int. Conf. Multimedia*, pp. 2823–2832, 2020.

[15] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR)*, pp. 5039–5049, 2021.

[16] B. K. Panigrahi, S. P. Mishra, and C. K. Samal, "Deepfake detection using deep learning: A review," *Advances in Research*, vol. 26, no. 4, pp. 555–564, 2025.

[17] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 15044–15054, 2021.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.