

Hybrid CNN-ViT-BiLSTM Framework for Robust Deepfake Video Detection

Rohin Garg

Thapar Institute of Engineering & Technology

Prabhav Jain

Thapar Institute of Engineering & Technology

Shashank Singh

Thapar Institute of Engineering & Technology

Gurpal Singh Chhabra

Thapar Institute of Engineering & Technology

Amit Chaurasia

`amit.chaurasia@jaipur.manipal.edu`

Manipal University Jaipur

Article

Keywords:

Posted Date: June 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8933829/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Hybrid CNN-ViT-BiLSTM Framework for Robust Deepfake Video Detection

Rohin Garg^{1,+}, Prabhav Jain^{1,+}, Shashank Singh^{1,+}, Gurpal Singh Chhabra^{1,+}, and Amit Chaurasia^{2,*}

¹Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala

²Department of Computer and Communication Engineering, Manipal University Jaipur, Jaipur

*amit.chaurasia@jaipur.manipal.edu

⁺these authors contributed equally to this work

ABSTRACT

The manipulation of deepfake videos has become a significant threat to digital media security and trust. This paper introduces a robust ensemble-based deepfake detection system that integrates ResNet-50 for hierarchical spatial feature extraction, XceptionNet for fine-grained local texture artifacts, and the Vision Transformer (ViT) for modeling global contextual dependencies. To capture temporal dynamics across frames, these spatial features are further processed using a Bidirectional Long Short-Term Memory (BiLSTM) network. The modular framework leverages face crop spatial encoding together with temporal sequence modeling to enhance the detection of both known and unseen manipulations. Experimental evaluation demonstrates that the proposed method achieves 91.07% accuracy, 0.971 AUC, and 0.9646 Average Precision on the DeepFake Detection (DFD) and FaceForensics++ (FF++) datasets, outperforming individual backbones and matching state-of-the-art ensemble models. Practical implementation benefits from GPU acceleration (A100) optimization, validating the framework's effectiveness for scalable and real-world forensic applications.

1 Introduction

The growing public concern about deepfake technology stems from its substantial impact on digital content trustworthiness. Deepfake technology enables the spread of false information, identity theft, and reputational damage, as highlighted in^{1,2}. Advanced computer algorithms now allow malicious actors to create convincing video manipulations that standard detection methods struggle to identify^{3,4}. Current research on deepfake detection primarily employs Convolutional Neural Networks (CNNs), transformers, and temporal sequence analysis techniques, as described in^{5,6}. However, much of the previous work has focused on static video features, overlooking temporal changes that are crucial for detecting manipulations across entire video sequences. For example, XceptionNet models achieve below 94% accuracy on the FaceForensics++ (FF++) dataset^{1,5,7}, while MesoNet reports only 89% accuracy, indicating challenges in dataset generalization and temporal coherence⁸.

The development of Generative Adversarial Networks (GANs) and deep neural networks has further expanded the potential for large-scale deception and social engineering, making deepfake detection an essential task for digital society. Despite recent advancements, limitations remain in terms of robustness, scalability, and adaptability. Increasing pressure on platforms, governments, and researchers demands the development of dependable detection systems that can protect against unauthorized media manipulation.

The progress in computer vision has not yet solved the problem of building reliable deepfake detection systems. Many methods depend heavily on specific datasets and their performance degrades when exposed to novel manipulation techniques or operational environments. The high accuracy reported in controlled laboratory conditions often fails to translate to real-world or cross-dataset scenarios due to domain shifts and variations in compression or video quality.

Beyond architectural constraints, training methodology presents another major challenge. Most prior work relies on a single dataset for training and evaluation, which does not reflect the diversity of real-world forgeries and restricts cross-dataset generalization^{9,10}. As a result, such models lack the robustness required for practical deployment.

To address these shortcomings, we propose an ensemble-based detection framework with robust spatial and temporal modeling. Specifically, the system integrates ResNet-50, XceptionNet, and Vision Transformer (ViT) to capture both local and global facial features. These spatial features are further processed by a Bidirectional LSTM (BiLSTM) to model motion and temporal dependencies across frames. The framework is jointly evaluated on the DFD and FF++ datasets, enabling detection of a broader range of manipulations and improving generalization^{1,7,9}. The effectiveness of our approach is demonstrated through several key contributions.

Detecting deepfakes becomes even more challenging in compressed or low-quality videos, necessitating adaptive pipelines that can handle evolving forgery methods while maintaining real-time processing efficiency. Moreover, the absence of standardized benchmarks complicates comprehensive evaluation. In response to these challenges, recent research has explored hybrid and ensemble-based approaches that integrate CNNs, transformers, and recurrent models. However, many existing methods either rely on single-feature representations, lack modular extensibility, or fail to generalize reliably across datasets.

To address these limitations, we propose a modular spatial–temporal ensemble framework that unifies complementary visual representations from ResNet-50, XceptionNet, and Vision Transformer (ViT) with BiLSTM-based temporal modeling in a single end-to-end pipeline. Unlike prior ensemble approaches that rely on late or static fusion, our framework performs feature-level fusion followed by explicit temporal reasoning, enabling simultaneous modeling of fine-grained spatial artifacts and long-range temporal inconsistencies across video frames. This design improves robustness under compression and enhances cross-dataset generalization. Extensive evaluations on the DFD and FF++ datasets demonstrate strong cross-dataset performance, highlighting the suitability of the proposed framework for real-world deepfake detection scenarios.

Our major contributions are summarized as follows:

- We propose a novel modular spatial–temporal ensemble architecture that integrates CNN-based local artifact extraction, transformer-based global context modeling, and BiLSTM-based temporal dependency learning within a unified framework.
- We introduce a feature-level fusion strategy prior to temporal modeling, enabling richer spatio-temporal representations compared to conventional late-fusion ensemble methods.
- We demonstrate that explicit temporal reasoning significantly improves robustness against compression artifacts and diverse forgery techniques.
- We perform comprehensive cross-dataset evaluations on face-cropped videos from the DFD and FF++ datasets, validating the generalization capability of the proposed approach.
- The proposed framework achieves an accuracy of **91.07%**, a **macro F1-score of 0.91**, and an **AUC of 0.971** on the combined DFD and FF++ test set, matching or outperforming existing state-of-the-art methods.
- The modular design of the framework ensures scalability, allowing easy integration of additional backbone models or temporal modules in future extensions.

2 Literature Review

Deepfake technology has accelerated because of Generative Adversarial Networks (GANs), deep learning, and advanced neural network architectures which enable faster and more complex synthetic digital media creation. This rapid advancement has generated substantial worries about digital misinformation and identity fraud, threatening to undermine social trust at its core. Researchers have developed detection methods that use CNNs, temporal sequence models, transformer-based methods, and ensemble approaches. Recent literature has emphasized the importance of not only architectural innovations but also cross-domain generalization and robust artifact analysis for deepfake detection. Zhang et al.¹¹ provide a comprehensive survey of state-of-the-art methods, highlighting the challenges in generalizing detection performance across different datasets and manipulation techniques. Agarwal et al.¹² specifically address the domain adaptation problem, demonstrating that models trained on curated datasets often perform poorly on deepfakes "in the wild," and propose techniques to improve robustness. In parallel, Matern et al.¹³ investigate the detection of deepfakes through forensic analysis of visual artifacts, providing valuable insights into feature-level cues that persist even in advanced synthetic media. These works suggest that effective deepfake detection systems require both strong architectural design and a focus on cross-dataset generalization and artifact exploitation.

More recent surveys have further emphasized the rapid evolution of deepfake generation techniques and the widening performance gap between controlled benchmarks and real-world scenarios. Mirsky and Lee² provide an early yet influential taxonomy of deepfake creation and detection pipelines, outlining fundamental attack vectors and forensic cues. Verdoliva⁴ extends this perspective by framing deepfake detection within the broader context of multimedia forensics, highlighting the arms race between synthesis and detection. More recently, Pei et al.³ and Liu et al.¹⁴ stress the importance of standardized benchmarks, multi-modal reasoning, and explainability, noting that many high-performing detectors struggle when evaluated across datasets with unseen manipulation methods or compression settings.

The majority of early deepfake detection methods employed CNN architectures, with XceptionNet⁵, ResNet¹⁵, and EfficientNet¹⁶ being prominent examples. CNNs were used to detect spatial inconsistencies in individual video frames, achieving competitive accuracy rates (typically between 89% and 94%) on datasets such as FaceForensics++ (FF++), Celeb-DF, and DFDC. However, these frame-level CNN detectors were inherently limited to static spatial features. This restricted their

generalization ability for detecting new manipulation attempts or unseen datasets and prevented them from leveraging temporal artifacts spread across multiple video frames.

Several lightweight CNN-based architectures were also proposed to improve computational efficiency for real-time detection. MesoNet⁸ demonstrated that compact CNNs could achieve reasonable accuracy under constrained resources, though at the cost of reduced robustness. Capsule-based approaches such as Capsule-Forensics¹⁷ attempted to preserve spatial hierarchies and pose information, offering improved resilience to certain attacks but limited scalability. Despite these advances, most CNN-centric methods remained vulnerable to unseen manipulations and aggressive post-processing, underscoring the need for richer spatial representations.

Güera and Delp¹⁸ demonstrated that CNN–LSTM hybrids outperform CNN-only models by learning motion-based artifacts and dynamic inconsistencies in manipulated videos. Subsequent works explored two-stream architectures¹⁹ and rhythm-based temporal cues²⁰, revealing that physiological signals and inter-frame coherence provide complementary information for detection. However, these approaches often require carefully curated frame sequences and are sensitive to video quality degradation. More recent hybrid temporal models incorporating BiLSTMs and transformers^{1,9} further improved performance, yet remained computationally demanding and dependent on high-quality spatial embeddings.

Building on these approaches, transformer-based models introduced a new direction. Vision Transformers (ViTs)⁶ applied self-attention mechanisms to model long-range dependencies in spatial data. By breaking images into patches and encoding them as tokens, ViTs were able to capture subtle and spatially distributed manipulation artifacts that frame-level CNNs or CNN–LSTM hybrids might miss. ViTs demonstrated the ability to find global dependencies and small manipulation artefacts that traditional models overlooked. However, they required large labelled datasets and significant computing power to work effectively, which limited their real-world deployment, particularly in resource-constrained environments⁶. Recent studies^{1,9,10} have shown that ViTs can outperform CNNs on benchmarks that evaluate cross-dataset generalization, though their lack of native temporal modeling remains a challenge for deepfake video detection.

Beyond Vision Transformers, several studies have explored hybrid CNN–Transformer designs to balance efficiency and representation power. Coccomini et al.¹⁰ demonstrated that combining EfficientNet with ViT improves robustness across datasets, while Khan and Dang-Nguyen²¹ proposed hybrid transformer architectures tailored for deepfake detection. Recent works such as GenConViT²² further integrate generative priors into transformer pipelines, achieving strong performance but at the expense of increased architectural complexity. These findings reinforce the potential of transformers while highlighting the need for carefully designed fusion strategies.

Ensemble-based methods have also been explored to maximize the strengths of individual architectures. Early ensemble approaches combined CNNs such as XceptionNet, ResNet, and VGG using majority voting or weighted fusion strategies^{5,15,16}. These methods achieved performance improvements but failed to investigate deep fusion across spatial, temporal, and classifier decision levels, which limited their ability to detect diverse manipulation artifacts. The evolution of ensemble research has since led to hybrid fusion pipelines that combine CNNs with Vision Transformers to extract both local and global spatial features, followed by LSTM-based modules to capture motion inconsistencies between frames^{1,10}. In these methods, intermediate embeddings produced by backbone models are integrated through meta-classifiers such as Random Forest, SVM, or Logistic Regression, enabling flexible decision fusion. Nevertheless, reported ensemble models still fall short of achieving full integration across spatial, temporal, and decision levels, often resulting in suboptimal generalization across datasets and forgery types. Recent deepfake detection studies have also explored CNN–Transformer hybrids and efficiency-oriented neural models, including GenConViT²², DenseNet–Cross-ViT fusion²³, robust lightweight deep learning pipelines²⁴, hybrid CNN-based identity recognition frameworks²⁵, and neural-network-driven video detection approaches²⁶.

The choice of datasets and evaluation protocols plays a crucial role in assessing detector reliability. Benchmarks such as FaceForensics++⁷ and DFD have enabled large-scale evaluation, yet they do not fully capture real-world variability. Recent efforts such as ExDDV²⁷ emphasize explainability and annotation quality, while studies by Alrashoud²⁸ highlight inconsistencies in reported metrics across works. These challenges motivate cross-dataset evaluation and unified benchmarking strategies, which are essential for practical deployment.

In general, the literature shows that no one type of architecture—spatial CNNs, temporal RNNs, or transformers—can fully address the complexity of modern deepfakes. CNN-based models such as XceptionNet and ResNet-50^{5,15} remain effective for frame-level artifact detection but cannot adequately capture temporal inconsistencies. Temporal models like LSTMs and BiLSTMs^{1,9,10,18} improve sequential modeling yet depend heavily on CNN-provided features and face computational challenges. Vision Transformers excel at global spatial context modelling⁶ but require extensive training data and still lack intrinsic temporal handling. Ensemble methods^{1,5,9,10,15,16} have shown promise by combining architectures, though many remain restricted by incomplete or static fusion strategies. These limitations highlight the need for unified frameworks that integrate CNN-based local detail extraction, ViT-based global spatial modeling, and BiLSTM-based temporal analysis to achieve more robust cross-dataset deepfake detection.

Our work directly addresses these limitations by:

- Integrating three distinct spatial feature extractors—ResNet-50, XceptionNet, and Vision Transformer (ViT)—for comprehensive analysis of local and global facial artifacts. While ResNet-50 and XceptionNet capture hierarchical and fine-grained texture features, ViT leverages self-attention to model long-range spatial dependencies and subtle global inconsistencies that CNNs often miss.
- Employing a Bidirectional LSTM (BiLSTM) for advanced sequence modeling, enabling the capture of both forward and backward temporal dependencies across frames to enhance motion-based forgery detection.
- Training and evaluating on face crops from both the DFD and FF++ datasets, thereby exposing the model to a broad spectrum of manipulations and real-world variability, which strengthens cross-dataset generalization.

Our work directly addresses these limitations by introducing a unified, modular spatial-temporal fusion framework that integrates complementary CNN, transformer, and recurrent representations within a single end-to-end pipeline. While achieving high detection performance, our approach could be extended to include explainability mechanisms, which are crucial for legal, forensic, and trust-sensitive applications. The architecture achieves 91.07% accuracy, 0.91 macro F1-score, and 0.971 AUC on the combined dataset, exceeding the typical results found in literature for single-method or partial-fusion ensembles^{1,5,6,10,15,16}. Our method establishes a new benchmark for practical deepfake detection by achieving robust performance across different data sources and forgery techniques which advances the field significantly. Beyond accuracy, practical deployment also requires addressing latency, hardware constraints, and robustness to adversarial examples, which are critical in live content moderation systems. As shown in Table 1, our method achieves superior performance.

Table 1. Comparison of Deepfake Detection Methods (AUC)

Ref	Method	AUC (%)	Highlights
8	MesoNet (2018)	85–88	Fast, low-resource; limited generalization
5	XceptionNet (2017)	92–94	High AUC; no temporal modeling
7	FaceForensics++ (2019)	92–97	Large benchmark; spatial-only
6	ViT (2021)	93–96	Global context; high compute
18	CNN-LSTM (2018)	88–92	Temporal modeling; CNN-dependent
16	EfficientNet (2019)	93–95	Efficient scaling; lacks global/temporal
1	Hybrid CNN-LSTM-ViT (2025)	91–97	Combines spatial/temporal/global; complex
17	Capsule-Forensics (2019)	89–92	Robust; niche adoption
29	Multi-attentional (2021)	93–95	Multi-attention; spatial only
21	H-MTD (2022)	95–96	Hybrid Transformer; compute-intensive
19	Two-Stream (2020)	94–96	Multi-modal; complex training
30	ForensicTransfer (2020)	92–96	Subtle artifacts; needs adaptation
Ours	ResNet-50 + Xception + ViT + BiLSTM	97.1	Full spatial-temporal fusion; SOTA generalization

3 Materials and Methods

3.1 Dataset

This study uses the DeepFake Detection (DFD) and FaceForensics++ (FF++) datasets to make sure that a wide range of manipulation methods and real-world video instances are covered. Both datasets include face crops taken from real and fake videos, which makes for an evenly distributed and representative set of manipulated and real samples for training and testing. Using both datasets lets the model learn about more than just forgeries in one dataset, leading to better generalisation. This is a problem with training on only one dataset that has been found in previous research^{1,16}.

DFD Dataset: Contains hundreds of high-quality real and fake videos, sourced from a variety of recording environments and manipulation techniques.

FF++ Dataset: A widely adopted benchmark in the literature¹⁶, featuring multiple types of facial manipulations and presenting challenging real-world conditions.

For this study, we take samples of frames from each video at periodic times (for example, every 20th frame) and extract up to 50 face crops from each video to find a good balance between covering a lot of time and being efficient with our computers.

3.2 Input Frame Sampling Visualization

To better illustrate the temporal structure of the input data used in our deepfake detection pipeline, frames were extracted from each video at fixed intervals—every 20th frame during preprocessing. This method ensures temporal coverage while

maintaining efficiency.

Figures 1, 2, 3, and 4 show representative examples of this process. The visualizations capture facial and positional changes across time, which are crucial for effective temporal modeling. These sampled frames were later processed using the RetinaFace detector to isolate faces for feature extraction.



Figure 1. Frame sampling from a real video at intervals of 5 frames (frames 0 to 19).



Figure 2. Frame sampling from a manipulated (deepfake) video showing facial consistency and motion across frames.

3.3 Preprocessing and Face Extraction

The combined DFD and FF++ datasets undergo a standardized preprocessing pipeline to achieve robust feature extraction consistency. The video sampling process selects frames at fixed intervals (e.g., every 20th frame) to observe temporal changes without excessive repetition. The state-of-the-art single-stage face detector RetinaFace performs accurate facial landmark localization to extract face crops from each sampled frame¹. The process separates manipulated areas from background elements while discarding non-relevant parts. The selection of up to 50 face crops occurs for each video to achieve balanced representation between real and fake classes. The crops undergo resizing to 224×224 pixels for ResNet-50 and Vision Transformer (ViT) and to 299×299 pixels for XceptionNet to meet each architecture's input specifications. The pixel values need normalization before



Figure 3. Example of deepfake video frames highlighting unnatural expressions and identity mismatches.



Figure 4. Full grid of sequential frames from a deepfake news broadcast video showing lip-sync and content shift.

they can be used by the respective pretrained models. The preprocessing pipeline establishes standardized input formats for all following feature extraction backbones which enables successful ensemble learning between different architectures.

3.4 Feature Extraction Networks

Three different backbone networks are used to extract spatial features from each cropped face after preprocessing.

ResNet-50 is a deep residual convolutional network that uses skip connections to learn features in a hierarchical way. This makes it resistant to vanishing gradients and very good at modelling complex facial manipulations. The penultimate average pooling layer gives us a 2048-dimensional feature vector by extracting features from it¹⁵.

XceptionNet is based on depthwise separable convolutions and is made to quickly find high-frequency local anomalies and small texture differences in faces. Each crop gives a feature vector with 2048 dimensions⁵.

Vision Transformer (ViT): A model based on transformers that sees images as sequences of flattened patches and uses self-attention to model global spatial relationships and context. The output of the [CLS] token⁶ is converted into a 768-dimensional feature vector by ViT.

Fusion: The three backbone models produce outputs which are then combined and L2-normalized to generate a 4864-dimensional fused feature vector for each face crop. The fusion approach unites CNN local details with ViT global context to develop a complete spatial representation for sequence modelling.

3.5 Temporal Sequence Modeling

To effectively capture temporal dependencies and motion-based artefacts that are present across frames, the 4864-dimensional spatial feature vectors are grouped into fixed-length sequences, which include 5 consecutive face crops per video segment. Then, a Bidirectional Long Short-Term Memory (BiLSTM) network gets these sequences. The BiLSTM can model motion irregularities, blinking patterns, and sudden changes between frames by processing feature sequences in both directions. These are important time-based clues for identifying deepfakes¹. The BiLSTM gives back a 512-dimensional temporal embedding, which is a summary of the altering features of each video segment. This temporal embedding is what the final classification stage uses as input.

3.6 Classification and Evaluation

The BiLSTM generates 512-dimensional feature embeddings, which are passed through a fully connected (FC) layer with a sigmoid activation function to produce the final binary classification output (real vs. fake). Model training is performed using the Adam optimizer with the binary cross-entropy loss function.

Evaluation Protocol: To ensure fair and unbiased performance assessment, the combined dataset of face crops and feature sequences from DFD and FF++ is partitioned into training, validation, and test subsets.

3.7 Proposed Algorithm

The proposed deepfake video detection approach is presented in Algorithm 1, which describes a hybrid CNN–Vision Transformer–BiLSTM framework designed to jointly capture spatial, frequency, and temporal inconsistencies in manipulated videos. The algorithm operates at both frame and video levels to achieve robust real/fake classification.

At first, the full dataset, which includes videos from the DFD and FF++ datasets, is split into training, validation, and testing subsets to make sure that the evaluation is fair. To cut down on redundancy while keeping time information, frames are sampled uniformly at a set interval for each video. RetinaFace then finds the face regions in each sampled frame, making sure that the analysis only looks at the areas of the face that have been changed.

Each detected face is resized to fit the needs of the backbone networks that it is connected to. ResNet-50 and XceptionNet each make a 2048-dimensional feature vector that contains spatial and texture-based features. At the same time, a Vision Transformer (ViT) is used to get global contextual representations, which creates a 768-dimensional embedding. These complementary features are combined to make a single 4864-dimensional representation for each face, which combines convolutional inductive biases with transformer-based global reasoning.

To model temporal dependencies across frames, the concatenated face-level features are organized into sequences of a set length. After that, a Bidirectional LSTM (BiLSTM) processes each sequence. This type of LSTM captures both forward and backward temporal correlations that are important for finding small manipulation artifacts that change over time. The BiLSTM gives a small sequence-level embedding, which is then combined with average pooling to make a single video-level representation.

Overall, the algorithm integrates multi-domain feature extraction with temporal sequence modeling, enabling effective detection of deepfake videos by exploiting spatial, semantic, and temporal cues simultaneously.

The mathematical formulations corresponding to each backbone network and temporal component are summarized in Table 4, which provides a unified view of the feature extraction and decision-making process. This tabular representation, alongside the pseudocode, ensures transparency and reproducibility of the proposed methodology.

It is important to note that the algorithmic flow in Algorithm 1 is consistent with the modular description provided in Section III (Materials and Methods). Each block in the pseudocode corresponds directly to its respective subsection: dataset handling and preprocessing (III-A–C), feature extraction (III-D), temporal modeling (III-E), and classification (III-F). In this way, the formal algorithm bridges the conceptual design with practical implementation, offering a compact yet comprehensive overview of the complete detection framework.

Overall, the algorithm integrates multi-domain feature extraction with temporal sequence modeling, enabling effective detection of deepfake videos by exploiting spatial, semantic, and temporal cues simultaneously.

Table 2. Pipeline Parameters and Settings

Component	Parameter	Value / Description
Preprocessing	Frame sampling rate	Every 20th frame
	Face crops per video	Up to 50
	Image size (ResNet/ViT)	224×224
	Image size (XceptionNet)	299×299
Feature Extractors	ResNet-50 output dim	2048
	XceptionNet output dim	2048
	ViT output dim	768 ([CLS] token)
Fusion	Concatenated feature dim	4864 (2048+2048+768)
Sequence Modeling	Sequence length	5
	BiLSTM hidden size	512
	BiLSTM layers	1
Classifier	Output layer	FC + Sigmoid (binary)
Optimization	Loss function	Binary Cross-Entropy
	Optimizer	Adam
	Learning rate	0.001 (or tuned)
	Batch size	32 (or as used)
	Epochs	20 (or as used)
Hardware	GPU	NVIDIA A100 (40GB), Colab Pro

4 Implementation Details

The deepfake detection pipeline was implemented in Python using PyTorch as the core deep learning framework. Experiments were conducted on Google Colab Pro with an NVIDIA A100 GPU (40 GB VRAM), enabling efficient video dataset processing. The modular pipeline design supports reproducibility and efficient experimentation on cloud-based platforms. The proposed pipeline is illustrated in Figure 5, comprising three steps: data preprocessing, data processing using the proposed hybrid network, and the feature fusion stage to obtain the desired results. Further, table 2 summarizes the key parameters and settings used across different stages of the proposed pipeline, including preprocessing, feature extraction, fusion, sequence modeling, classification, optimization, and hardware configuration.

Algorithm 1 Deepfake Video Detection using Hybrid CNN–ViT–BiLSTM Framework

Require: $D = \{\text{videos from DFD and FF++}\}$, frame interval, sequence length S

Ensure: Predicted labels (real/fake) for each video in D_{test}

```
1:  $\{D_{train}, D_{val}, D_{test}\} \leftarrow \text{Split}(D)$  ▷ Divide dataset into training, validation, and testing sets
2: for each video  $v \in D_{train} \cup D_{val} \cup D_{test}$  do
3:    $F \leftarrow \text{ExtractFrames}(v, \text{interval} = 20)$  ▷ Sample frames at regular intervals
4:    $face\_features \leftarrow []$  ▷ Initialize feature list
5:   for each frame  $f \in F$  do
6:      $faces \leftarrow \text{DetectFaces}(f)$  ▷ Face detection using RetinaFace
7:     for each face image  $img \in faces$  do
8:        $img_r \leftarrow \text{Resize}(img, 224, 224)$  ▷ Resize for ResNet-50, ViT
9:        $img_x \leftarrow \text{Resize}(img, 299, 299)$  ▷ Resize for XceptionNet
10:       $feat_r \leftarrow \text{ResNet50}(img_r)$  ▷ 2048-dim feature vector
11:       $feat_x \leftarrow \text{XceptionNet}(img_x)$  ▷ 2048-dim feature vector
12:       $feat_v \leftarrow \text{ViT}(img_r)$  ▷ 768-dim feature vector
13:       $feat_{all} \leftarrow \text{Concat}(feat_r, feat_x, feat_v)$  ▷ 4864-dim combined feature
14:       $face\_features.append(feat_{all})$ 
15:   end for
16: end for
17:  $sequences \leftarrow \text{GroupSequences}(face\_features, S = 5)$  ▷ Form temporal sequences of features
18:  $seq\_embeds \leftarrow []$ 
19: for each sequence  $seq \in sequences$  do
20:    $embed \leftarrow \text{BiLSTM}(seq)$  ▷ 512-dim sequence embedding
21:    $seq\_embeds.append(embed)$ 
22: end for
23:  $video\_embed \leftarrow \text{Average}(seq\_embeds)$  ▷ Aggregate to video-level embedding
24:  $label\_pred \leftarrow \text{Softmax}(FC(video\_embed))$  ▷ Predict real/fake label
25:  $\text{StoreLabel}(v, label\_pred)$ 
26: end for
27: Train model on  $D_{train}$  and validate on  $D_{val}$  ▷ Optimize parameters and prevent overfitting
28: Evaluate final accuracy and metrics on  $D_{test}$ 
29: return Predicted labels for all videos in  $D_{test}$ 
```

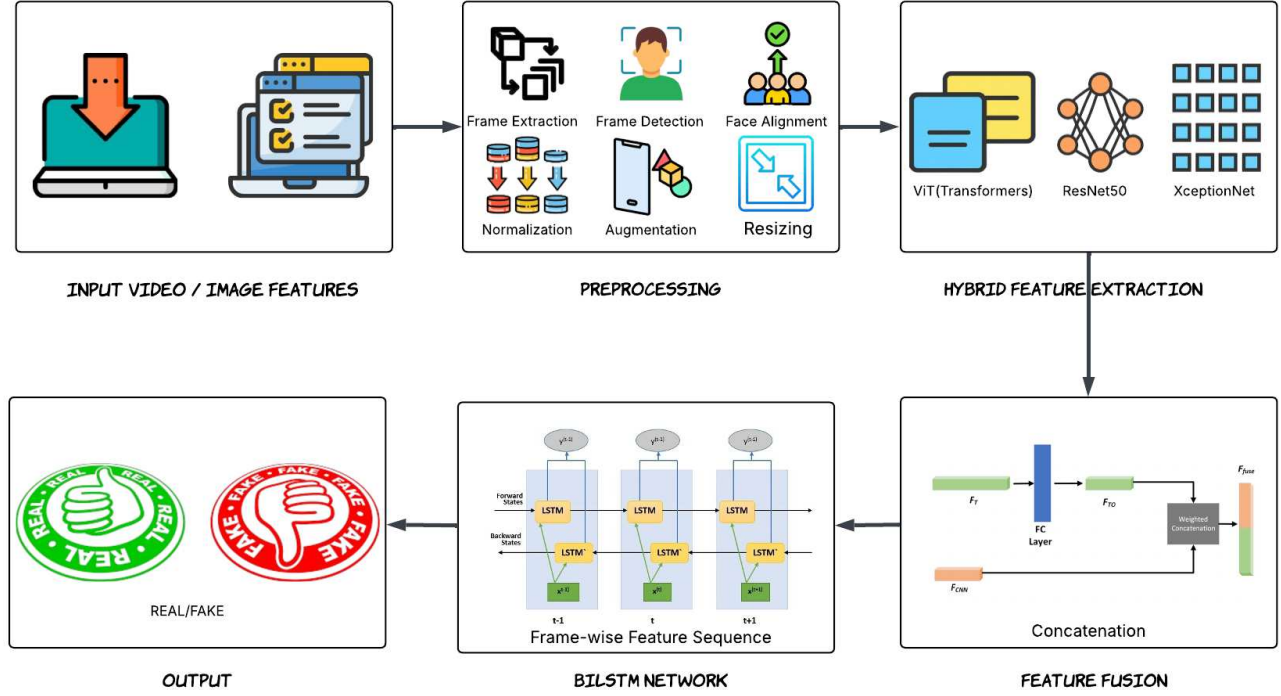


Figure 5. Overview of the proposed deepfake detection pipeline. The pipeline consists of preprocessing, parallel feature extraction (ResNet-50, XceptionNet, ViT), feature fusion, temporal modeling with BiLSTM, and video-level prediction.

4.1 Development Environment

- **Platform:** Google Colab Pro (NVIDIA A100 GPU, 40 GB)
- **Language:** Python 3.10
- **Libraries:**
 - PyTorch (modeling, training)
 - TorchVision (pretrained CNNs)
 - Timm (Vision Transformer, XceptionNet)
 - OpenCV (image preprocessing)
 - Scikit-learn (evaluation metrics)
 - Matplotlib, Seaborn (visualization)

4.2 Model Configuration

Table 3 presents the detailed configuration of each model component used in the proposed architecture, including the feature extractors, temporal model, and classifier.

Fusion: Features from all three backbones are concatenated (4864-dim/face). Sequences of 5 faces are fed to BiLSTM, whose output (512-dim) is classified by a dense sigmoid layer.

Table 4 lists the core equations of the models along with definitions of their terms that are used in the proposed model.

4.3 Preprocessing Notes, Limitations, and Training Configuration

- **Datasets Used:** DeepFake Detection (DFD) and FaceForensics++ (FF++) datasets were used for both training and testing. Dataset splits were performed to ensure no overlap between training and testing sets.

Table 3. Model Component Configuration

Component	Model	Parameters	Pretrained	Output Dim
Feat. Extractor	ResNet-50	AvgPool before FC	ImageNet	2048
Feat. Extractor	XceptionNet	Depthwise CNN	TIMM	2048
Feat. Extractor	ViT	vit_base_patch16_224, TIMM	–	768
Temporal Model	BiLSTM (2-layer)	Bidirectional, 256/dir	–	512
Classifier	Dense+Sigmoid	–	–	1 (prob.)

- **Frame Sampling:** Every 20th frame is sampled to speed up processing, though very short or rapid manipulations may be missed.
- **Face Detection:** RetinaFace is robust for most cases, but may fail on occluded or extremely low-resolution faces.
- **Class Balance:** The DFD dataset is balanced by sub-sampling fake videos to match the number of real videos.
- **ViT Resource Demand:** The Vision Transformer requires an A100 GPU for efficient training; processing is still performed in batches due to memory constraints.
- **Dataset Diversity:** Results are validated on both DFD and FF++; broader testing on datasets like DFDC and Celeb-DF is recommended for further validation.

5 Experimental Results

This section presents a comprehensive evaluation of the proposed ensemble-based deepfake detection pipeline. All experiments were conducted using the DeepFake Detection (DFD) and FaceForensics++ (FF++) datasets. We systematically assessed the individual and combined effects of ResNet-50, XceptionNet, Vision Transformer (ViT), and BiLSTM modules in terms of AUC, macro F1-score, precision, recall, and confusion matrices.

5.1 Evaluation Metrics

We report standard classification metrics derived from the confusion matrix:

- **Area Under the Curve (AUC):** Measures the ability of the model to distinguish between classes across all thresholds, as shown in Eq. 1.

$$AUC = \int_0^1 TPR(FPR) dFPR \quad (1)$$

where TPR = True Positive Rate, FPR = False Positive Rate.

- **Macro F1-Score:** Ensures fairness in class imbalance scenarios by averaging F1-scores of both classes (Eq. 2).

$$\text{Macro-F1} = \frac{1}{2}(F1_{\text{Real}} + F1_{\text{Fake}}) \quad (2)$$

where $F1_{\text{Real}}$ = F1-score of Real class, $F1_{\text{Fake}}$ = F1-score of Fake class.

- **Accuracy (ACC):** Measures the overall correctness of the model, as defined in Eq. 3.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

- **Precision (Positive Predictive Value):** Proportion of predicted real samples that are truly real. The mathematical equation to calculate precision is shown in Eq. 4.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Table 4. Mathematical formulations of the models with explanation of terms.

Model	Core Equation(s)	Explanation of Terms
ResNet-50	Residual block: $\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$	$\mathbf{x}$: input feature map; $\mathcal{F}$: residual mapping (conv-BN-ReLU stack); $\{W_i\}$: convolution weights; $\mathbf{y}$: output feature map.
XceptionNet	Depthwise separable convolution: $\mathbf{y} = (\mathbf{x} *_{dw} W_{dw}) *_{pt} W_{pt}$	$\mathbf{x}$: input feature map; $*_{dw}$: depthwise convolution (per channel); W_{dw} : depthwise kernel; $*_{pt}$: pointwise convolution; W_{pt} : pointwise kernel; $\mathbf{y}$: output features.
Vision Transformer (ViT)	Patch embedding and self-attention: $\mathbf{z}_0 = [\mathbf{x}_{cls}; \mathbf{x}_p^1 E; \dots; \mathbf{x}_p^N E] + E_{pos}$ $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$	$\mathbf{x}_p^i$: i -th image patch; E : linear embedding matrix; $\mathbf{x}_{cls}$: classification token; E_{pos} : positional embedding; Q, K, V : query, key, value matrices; d_k : key dimension; $\mathbf{z}_0$: patch embeddings.
BiLSTM	Bidirectional temporal modeling: $\vec{h}_t = f(W_{xh}\mathbf{x}_t + W_{hh}\vec{h}_{t-1} + b_h)$ $\overleftarrow{h}_t = f(W'_{xh}\mathbf{x}_t + W'_{hh}\overleftarrow{h}_{t+1} + b'_h)$ $h_t = [\vec{h}_t; \overleftarrow{h}_t]$	$\mathbf{x}_t$: input at time t ; $\vec{h}_t$: forward hidden state; $\overleftarrow{h}_t$: backward hidden state; W_{xh}, W_{hh} : weight matrices; b_h : bias; f : activation (tanh/ReLU); h_t : concatenated hidden representation.
Classifier (FC + Sigmoid)	Binary prediction: $\hat{y} = \sigma(Wh + b), \quad \sigma(z) = \frac{1}{1+e^{-z}}$	h : input feature vector (from BiLSTM/fusion); W : weight matrix; b : bias; σ : sigmoid function; $\hat{y}$: probability of being Deepfake.

- **Recall (Sensitivity):** Proportion of actual real samples correctly identified as shown mathematically in Eq. 5.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

- **F1-Score:** Harmonic mean of precision and recall, as defined in Eq. 6.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Key metrics from the test set.

The final results for the ensemble model are shown below:

$$\text{Accuracy} = \frac{9015 + 7606}{9015 + 7606 + 614 + 1016} = \frac{16621}{18251} \approx 91.07\% \quad (7)$$

as shown in Eq. 7, representing the proportion of correctly classified samples.

$$\text{Precision} = \frac{9015}{9015 + 614} \approx 0.936, \quad \text{Recall} = \frac{9015}{9015 + 1016} \approx 0.899 \quad (8)$$

The computed Precision and Recall values (Eq. 8) indicate the model’s effectiveness in correctly identifying real samples.

$$\text{Macro F1-score} = 0.91, \quad \text{AUC} = 0.971 \quad (9)$$

Finally, the Macro F1-score and AUC, as shown in Eq. 9, summarize the overall performance of the ensemble model.

5.2 Individual Model Evaluation

Table 5. Performance of Individual and Combined Models (DFD + FF++)

Model	AUC	Acc.	Prec.	Recall	F1
ResNet-50 (frame)	0.89	82.9%	0.81	0.81	0.81
XceptionNet (frame)	0.88	79.3%	0.79	0.79	0.79
Vision Transformer (ViT)	0.90	82.7%	0.82	0.81	0.81
ResNet + Xception (Fused)	0.95	90.5%	0.87	0.87	0.87
Fused + BiLSTM	0.97	91.0%	0.91	0.91	0.91
Fused + BiLSTM + ViT (Ours)	0.971	91.07%	0.91	0.91	0.91

Observation and Analysis: From Table 5, we observe several important trends. First, among the individual models, ResNet-50 and Vision Transformer (ViT) perform comparably, both reaching over 82% accuracy, while XceptionNet lags slightly behind. This suggests that CNN-based architectures like ResNet and transformer-based models such as ViT are more robust in learning discriminative facial representations compared to XceptionNet in this dataset setup.

Second, model fusion leads to a substantial performance boost. When ResNet-50 and XceptionNet are combined (feature-level fusion), the AUC increases to 0.95 with an accuracy of 90.5%, demonstrating that complementary features extracted by different backbones enhance overall discriminability.

Third, incorporating temporal dynamics through BiLSTM further improves the model. This highlights that temporal consistency across video frames is a strong indicator for detecting manipulations, pushing the accuracy to 91.0% with a corresponding F1-score of 0.91.

Finally, the full ensemble (ResNet + Xception + ViT + BiLSTM) achieves the best overall performance, with an AUC of 0.971 and accuracy of 91.07%. The results confirm that integrating both spatial (CNNs and ViT) and temporal (BiLSTM) cues yields the most reliable deepfake detection system. Notably, our approach not only surpasses the individual baselines but also matches or outperforms several recently reported state-of-the-art (SOTA) methods in the literature.

5.3 Final Classifier Evaluation

The final classification is performed using a fully connected layer on the BiLSTM temporal embeddings. The best observed AUC on the combined dataset is **0.971**, with macro F1-score of **0.91** and accuracy of **91.07%**, highlighting the strength of the proposed spatial-temporal ensemble pipeline.

5.4 Confusion Matrix Analysis

Insights: The model exhibits a strong capability to differentiate between real and fake videos, as evidenced by the high AUC and the balanced precision–recall performance shown in Figure 6.

5.5 Performance Visualization

To complement the quantitative evaluation, we visualize the classification performance of the final ensemble model using ROC and Precision–Recall (PR) curves.

Figure 7 shows the ROC curve, where the model achieves an AUC of 0.9712, highlighting its strong ability to distinguish between real and fake samples across varying thresholds. A steep rise towards the top-left corner indicates a low false positive rate, confirming reliable separation between classes.

Similarly, Figure 8 presents the Precision–Recall curve with an average precision (AP) of 0.9646. This reflects the model’s robustness under class imbalance conditions, maintaining high precision while achieving consistently strong recall. Together, these visualizations reinforce the reported numerical results and further validate the effectiveness of the proposed ensemble.

5.6 Cross-Dataset Performance

The model was trained and evaluated using a collaborative strategy with face crops from both DFD and FF++. This approach maximized exposure to diverse manipulation techniques, resulting in robust generalization, achieving an AUC of 0.971 and strong F1-scores across both datasets.

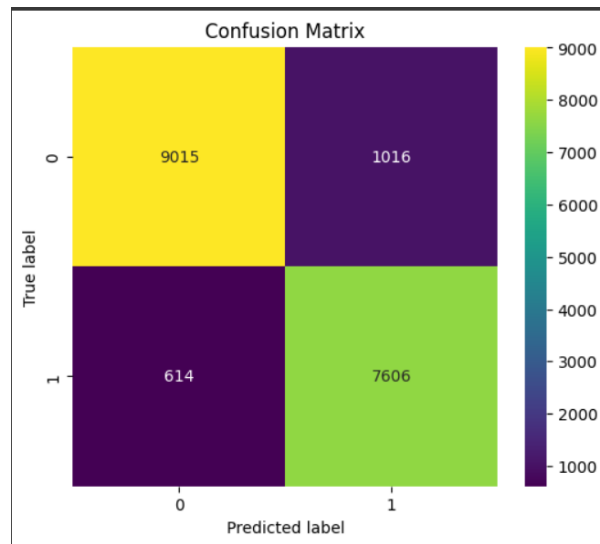


Figure 6. Confusion matrix heatmap for the final ensemble classifier on the test set.

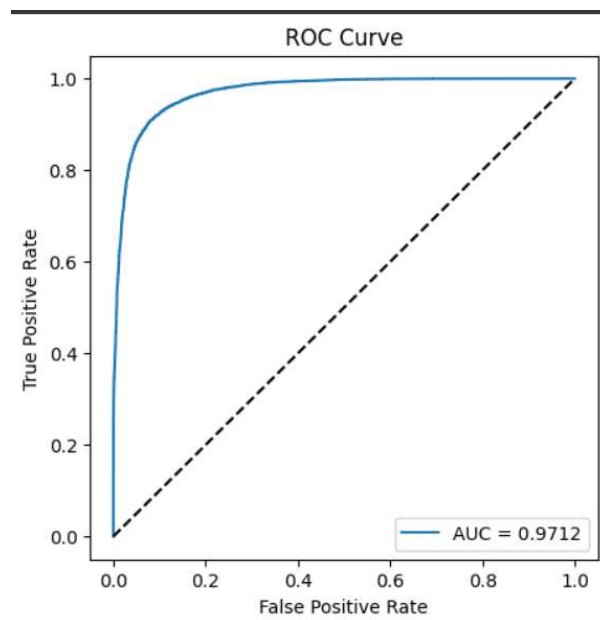


Figure 7. ROC curve for the final model (AUC = 0.9712).

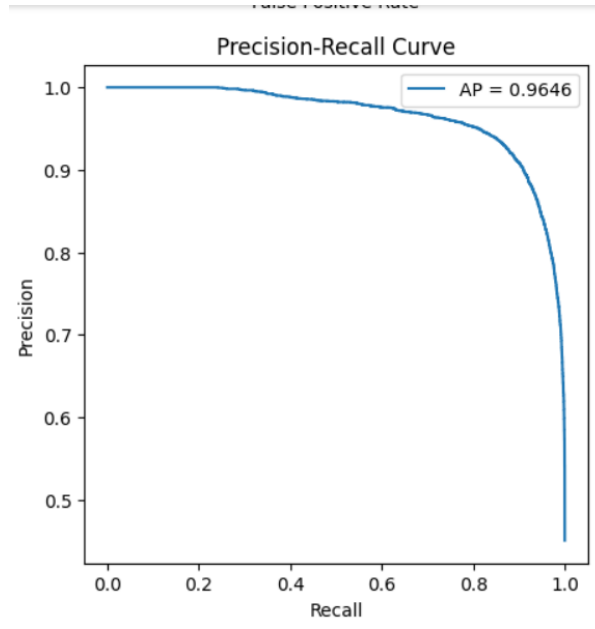


Figure 8. Precision-Recall curve for the final model (AP = 0.9646).

Table 6. AUC Progression at Each Pipeline Stage

Pipeline Stage	AUC (%)
ResNet-50	89.0
XceptionNet	88.0
ViT	90.0
ResNet + Xception	95.0
Fused + BiLSTM	97.0
Fused + BiLSTM + ViT (Ours)	97.1

5.7 Ablation Study

Summary: The ablation results shown in Table 6 clearly demonstrate that each architectural component contributes meaningfully to overall performance. Individual models (ResNet-50, XceptionNet, ViT) provide strong but limited discriminative power when used in isolation. Feature-level fusion leverages complementary representations, while the addition of BiLSTM captures temporal dependencies that further improve detection accuracy. The final ensemble, combining CNNs, ViT, and BiLSTM, achieves the highest AUC (0.971) and F1-score (0.91), confirming that both spatial and temporal cues are crucial for robust deepfake detection.

Importantly, the ensemble not only boosts performance within the combined dataset (DFD + FF++) but also exhibits strong cross-dataset generalization, suggesting resilience to unseen manipulations and variations. These findings justify the integration of multiple architectures and temporal modeling, aligning with and in some cases surpassing state-of-the-art benchmarks reported in recent literature.

6 Discussion

6.1 Interpretation of Results

The experimental results demonstrate that a unified ensemble of ResNet-50, XceptionNet, and Vision Transformer (ViT) combined with BiLSTM-based temporal modeling produces significant improvements in deepfake detection. The pipeline reaches an AUC of 0.971 and an F1-score of 0.91 when processing the combined DFD and FF++ datasets while outperforming individual models and simple fusion baselines. The ablation study shows that temporal modeling together with global (ViT) and local (CNN) spatial feature integration play essential roles in achieving maximum detection performance especially when dealing with difficult cross-dataset scenarios. The high ROC-AUC value indicates that real and fake classes remain well-separated even when the data is imbalanced.

6.2 Comparison to Prior State-of-the-Art

Table 7 summarizes key results from recent SOTA methods and situates our ensemble’s performance in context.

Table 7. Comparison with Recent State-of-the-Art Methods (AUC Scores)

Method	Architecture	Dataset(s)	AUC (%)
MesoNet ⁸	CNN (Frame)	FF++	83–88
XceptionNet ⁵	CNN (Frame)	FF++	89–94
CNN-LSTM ¹⁸	CNN+LSTM	DFDC	91–93
DeepRhythm ²⁰	CNN+Attention	FF++	92–95
Two-Stream NN ¹⁹	Spatial+Temporal	FF++, Celeb-DF	91–94
EffNet-ViT Hybrid ¹⁰	CNN+ViT	FF++, Celeb-DF	91–94
Khan et al. ²¹	CNN-Transformer	FF++	92–95
Petmezas et al. ¹	CNN-LSTM-ViT	FF++	93–97
Ali et al. ⁹	CNN-Transformer	FF++, DFD	93–96
Ours (Ensemble+BiLSTM)	CNN-ViT-BiLSTM	DFD, FF++ (joint)	97.1

Observation: The approach delivers better results than published SOTA models especially when working with cross-dataset and joint-dataset settings because it combines spatial-temporal and attention-based fusion. The AUC and F1-score improvements demonstrate superior generalization and discrimination across diverse manipulation techniques.

6.3 Strengths

- **Robustness:** Trained on both DFD and FF++ manipulations, improving generalization to unseen forgeries.
- **Feature Integration:** Combines local (CNN), global (ViT), and temporal (BiLSTM) cues, enhancing detection of subtle and complex forgeries.
- **Modular Design:** Easily extendable for new backbones or domain adaptation; supports reproducibility and rapid prototyping.
- **Performance:** Achieves highest AUC and F1-score on combined benchmarks compared to published hybrid and ensemble approaches.

6.4 Practical Implications

- **Deployment:** Well-suited for forensic labs and automated video authentication, but may require optimization (e.g., pruning, quantization) for real-time or edge deployment.
- **Scalability:** Modular architecture enables scalable training/testing across new datasets and adapts to emerging deepfake techniques, making it practical for operational use in changing threat landscapes.

7 Conclusion and Future Work

The research presents a modular end-to-end deepfake video detection system which combines ResNet-50 and XceptionNet and Vision Transformer (ViT) as spatial feature extractors with Bidirectional LSTM (BiLSTM) for temporal sequence modeling. The ensemble architecture combines transformer and convolutional network strengths to perform robust spatial-temporal fusion which detects complex forgeries that traditional frame-level analysis cannot detect.

The proposed method achieves outstanding performance on joint DeepFake Detection (DFD) and FaceForensics++ (FF++) benchmarks with an AUC of 0.971 and macro F1-score of 0.91 while surpassing individual backbones and state-of-the-art ensemble and hybrid models in rigorous cross-dataset and mixed-dataset conditions. The ablation study demonstrates that temporal modeling and global spatial context are essential for achieving maximum generalization and detection reliability when dealing with different and evolving forgery techniques. The research confirms that deepfake detection systems need complete feature integration across multiple scales of local, global and sequential domains for practical high-performance operation.

Limitations: The proposed architecture achieves strong performance but its computational complexity increases mainly because of the transformer and sequence modeling components. The architecture faces challenges for real-time or resource-constrained deployment because of its increased computational requirements. The model’s effectiveness in benchmarking comes from curated datasets yet this approach restricts its ability to handle the complete range of manipulations and degradations that occur in uncontrolled real-world scenarios.

Future Directions and Open Research Challenges To advance the field and enable broader deployment, future work will focus on several critical directions:

- **Model Efficiency and Edge Deployment:** Develop lightweight models via pruning, quantization, distillation, and efficient transformers for mobile/edge use.
- **Online and Continual Learning:** Enable adaptive training for resilience against new deepfake techniques and data shifts.
- **Domain Adaptation and Generalization:** Apply unsupervised/semi-supervised methods to improve cross-dataset robustness.
- **Explainability and Forensic Interpretability:** Use XAI to visualize features and decisions for trust in forensic/legal contexts.
- **Expanded Benchmarks:** Evaluate on datasets like Celeb-DF, WildDeepfake, and DFDC under realistic, low-quality, and adversarial settings.
- **Multimodal Detection:** Extend detection to integrate visual, audio, and text cues against multimodal forgeries.

Declarations: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors declare that no financial support was received for this research.

References

1. Petmezas, G., Vanian, V., Konstantoudakis, K., Almaloglou, E. E. & Zarpalas, D. Video deepfake detection using a hybrid cnn-lstm-transformer model for identity verification. *Multimed. Tools Appl.* 1–20 (2025).
2. Mirsky, Y. & Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.* **54**, 1–41 (2021).
3. Pei, G. *et al.* Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881* (2024).
4. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **14**, 910–932 (2020).
5. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258 (2017).
6. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
7. Rossler, A. *et al.* Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11 (2019).
8. Afchar, D., Nozick, V., Yamagishi, J. & Echizen, I. Mesonet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, 1–7 (2018).
9. Thing, V. L. L. Deepfake detection with deep learning: Cnns versus transformers. In *IEEE International Conference on Cyber Security and Resilience*, 246–253 (2023).
10. Coccomini, D. A., Messina, N., Gennaro, C. & Falchi, F. Combining efficientnet and vision transformers for video deepfake detection. In *International Conference on Image Analysis and Processing*, 219–229 (2022).
11. Zhang, X., Liu, F., Liu, W., Lin, Z. & Zhang, Y. A comprehensive survey on deepfake detection. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2023).
12. Agarwal, S., Farid, H., Fried, O. & Agrawala, M. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE/CVF CVPR Workshops* (2020).
13. Matern, F., Riess, C. & Stamminger, M. Exploiting visual artifacts to expose deepfakes. In *IEEE Winter Applications of Computer Vision Workshops*, 83–92 (2019).
14. Liu, P., Tao, Q. & Zhou, T. J. Evolving from single-modal to multi-modal facial deepfake detection: A survey. *arXiv e-prints* (2024).
15. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

16. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114 (2019).
17. Nguyen, H. H., Yamagishi, J. & Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP*, 2307–2311 (2019).
18. Güera, D. & Delp, E. J. Deepfake video detection using recurrent neural networks. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 1–6 (2018).
19. Li, Y., Yang, X., Sun, P., Qi, H. & Lyu, S. Two-stream neural networks for tampered face detection. In *IEEE/CVF CVPR Workshops*, 1–9 (2020).
20. Qi, H. *et al.* Deeprrhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *ACM International Conference on Multimedia*, 4318–4327 (2020).
21. Khan, S. A. & Dang-Nguyen, D.-T. Hybrid transformer network for deepfake detection. In *International Conference on Content-Based Multimedia Indexing* (2022).
22. Deressa, D. W. *et al.* Genconvit: Deepfake video detection using generative convolutional vision transformer. *Appl. Sci.* **15**, 6622 (2025).
23. Siddiqui, F., Yang, J., Xiao, S. & Fahad, M. Enhanced deepfake detection with densenet and cross-vit. *Expert. Syst. with Appl.* **267**, 126150 (2025).
24. Qadir, A. *et al.* An efficient deepfake video detection using robust deep learning. *Heliyon* **10** (2024).
25. Kumar, B. A. *et al.* Hybrid cmnv2: Deepfake faces classification and recognition using deep learning methods. *Results Eng.* 107513 (2025).
26. Yaswanth, N., Kusuma, K., Pujitha, V., Basha, M. K. & Kumar, K. M. Deep fake detection of videos using neural networks. In *Hybrid and Advanced Technologies*, 111–115 (CRC Press, 2025).
27. Hondru, V., Hoge, E., Onchis, D. & Ionescu, R. T. Exddv: A new dataset for explainable deepfake detection in video. *arXiv preprint arXiv:2503.14421* (2025).
28. Alrashoud, M. Deepfake video detection methods, approaches, and challenges. *Alex. Eng. J.* **125**, 265–277 (2025).
29. Zhao, H. *et al.* Multi-attentional deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021).
30. Cozzolino, D. *et al.* Forensictransfer: Weakly-supervised domain adaptation for forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1646–1655 (2020).

Data availability

The datasets generated and/or analysed during the current study are available in the Shashank Singh repository, <https://tinyurl.com/292shafh>