GMR
The Original

EXPDF: Adversarially Robust Spatiotemporal Deepfake Detection Using Temporal-Aware Attention Fusion And Evolutionary Hyperparameter Optimization

Kanchan Warkar[1*], Rishikesh Rawat[2], Ghizal F. Ansari[3] and Sudhir Mohod[4]

[1*,2]Department of CSE, Madhyanchal Professional University, Bhopal, MP, India
[3]Department of Physics, Madhyanchal Professional University, Bhopal, MP, India
[4]Department of CSE, Bapurao Deshmukh College of Engineering, Sevagram, MH, India
*Corresponding Author(s): Kanchan Warkar, Email: kanchan22.warkar@gmail.com.
Contributing Author(s) Email(s): rawatrishikesh@gmail.com, gfansari@mpu.ac.in, sudhir_mohod@rediffmail.com.

Abstract. Deepfake videos generated using advanced artificial intelligence techniques have become increasingly difficult to distinguish from authentic content, creating major challenges for digital media authentication as well as multimedia forensics. Although recent progress has improved deepfake detection, many existing approaches still show limited cross-domain generalization, vulnerability to adversarial perturbations, and reduced reliability on unseen datasets. In addition, several detectors depend heavily on dataset-specific artifacts, which cause performance degradation under real-world conditions and raise concerns regarding reproducibility as well as practical deployment. To address these limitations, this study proposes an adversarially robust spatiotemporal deepfake detection framework that integrates Xception-based spatial feature extraction, bidirectional Long Short-Term Memory (BiLSTM) temporal modeling, and a Temporal-Aware Attention Fusion (TAF) module for adaptive feature aggregation. Robustness is improved through Fast Gradient Sign Method (FGSM), and Projected Gradient Descent (PGD) adversarial training, while the hybrid Marine Predators Algorithm–Genetic Algorithm (MPA–GA) strategy is used for automated hyperparameter optimization. The framework was evaluated on four benchmark datasets, including FaceForensics++ (FF++), Celeb-DF v2, DFDC, and WildDeepfake. Experimental results obtained from ten independent runs demonstrated accuracies of 98.2 ± 0.3%, 97.1 ± 0.4%, 95.8 ± 0.5%, and 93.9 ± 0.7% on FF++, Celeb-DF v2, DFDC, and WildDeepfake, respectively. During cross-dataset evaluation, the framework achieved 91.6 ± 0.6% accuracy when trained on FF++ and tested on Celeb-DF v2, which indicates strong transferability across unseen distributions. Under PGD adversarial attacks, the proposed model retained 86.2 ± 0.8% accuracy, demonstrating improved perturbation resilience. Additionally, Grad-CAM as well as SHAP analyses provided interpretable forensic evidence supporting model decisions. The proposed framework establishes a reproducible, robust, and explainable deepfake detection pipeline suitable for practical multimedia forensic applications.

Keywords: Adversarial training, Deepfake detection, Explainable artificial intelligence, Hyperparameter optimization, Spatiotemporal learning, Temporal attention fusion.

1 INTRODUCTION

Recent advances in generative artificial intelligence as well as deep learning have significantly improved the quality of synthetic multimedia generation. Deepfake videos produced using generative adversarial networks, autoencoders, and diffusion-assisted manipulation techniques are now capable of generating highly realistic facial expressions, lip synchronization, and identity transformations. Publicly available software tools together with large-scale video datasets have accelerated the development and distribution of manipulated media across social platforms, entertainment applications, and online communication systems [1]. As result, reliable deepfake detection has become important research problem in multimedia forensics as well as digital media authentication.

Existing deepfake detection approaches mainly rely on convolutional neural networks, recurrent neural networks, and transformer-based architectures to identify spatial artifacts and temporal inconsistencies in manipulated videos. CNN-based models are effective in learning local forgery traces, while temporal architectures such as LSTM networks improve detection of frame-level inconsistencies across video sequences. Although these methods have reported high classification accuracy on benchmark datasets, the performance often decreases under

real-world conditions involving video compression, blur, frame loss, and unseen manipulation distributions [2][3]. This limitation restricts the practical deployment of many current detection systems.

Another important challenge is the robustness as well as interpretability of deepfake detectors. Adversarial perturbations can alter input distributions and significantly reduce detection accuracy, even in high-performing architectures. Furthermore, many existing models operate as black-box systems and provide limited forensic evidence to explain prediction decisions. In multimedia forensic applications, interpretable outputs are important for validating manipulated regions, and improving user confidence in automated detection systems. Therefore, robust and interpretable deepfake detection frameworks with improved cross-domain generalization remain active research area [4].

Despite recent progress in deepfake detection, several practical limitations continue to affect deployment reliability. Deepfake detectors trained on benchmark datasets frequently experience performance degradation when exposed to compressed or low-quality videos. Models evaluated on unseen datasets also demonstrate limited generalization because of overfitting to dataset-specific artifacts. In addition, adversarial perturbations can significantly reduce classifier stability, and increase false predictions [5]. Another concern is the lack of transparency in deep learning-based forensic systems, where prediction decisions are often difficult to interpret. These challenges are summarized as follows:

Practical Issue	Current Failure
Compressed videos	Accuracy collapse
Unseen datasets	Poor generalization
Adversarial perturbation	Vulnerability
Black-box outputs	Low forensic trust

The aim of this study is to develop a robust and interpretable spatiotemporal deepfake detection framework with improved cross-domain generalization for multimedia forensic applications.

The major objectives of this study are as follows:

1. To learn spatial forgery representations using the Xception-based feature extraction framework.
2. To capture temporal inconsistencies across the video frames using LSTM-based sequence modeling.
3. To improve adversarial robustness through FGSM, and PGD-based adversarial training strategies.
4. To optimize model hyperparameters using the hybrid Marine Predators Algorithm-Genetic Algorithm (MPA–GA) optimization framework.
5. To provide interpretable forensic evidence using the explainable artificial intelligence techniques.

This study focuses only on video-based deepfake detection using RGB video frames. The framework is designed for facial manipulation detection in benchmark multimedia forensic datasets. Audio deepfake detection, and multimodal audio-visual synchronization analysis are outside the scope of this work. Similarly, the proposed system does not address fully synthetic diffusion-generated videos or text-to-video generation frameworks.

Although recent deepfake detection studies have reported strong benchmark performance, several limitations still remain unresolved. Existing models often show reduced cross-domain robustness because of overfitting to dataset-specific visual artifacts [6][7]. Adversarial resilience also remains limited, particularly under perturbation-based attacks such as FGSM as well as PGD. In optimization strategies, most current methods use fixed hyperparameter configurations or isolated tuning approaches without joint optimization of temporal, robustness, and fusion parameters. Furthermore, explainability in deepfake detection is still largely dependent on posthoc visualization methods without systematic forensic interpretation (see Table 1).

Table 1. Summary of Research Gap

Research Gap	Existing Limitation
Cross-domain robustness	Dataset overfitting
Adversarial resilience	Weak attack robustness
Optimization	Limited joint hyperparameter search
Interpretability	Mostly posthoc

The main contributions of this study are summarized as follows:

1. A spatiotemporal fusion framework integrating Xception-based spatial learning, and LSTM-based temporal modeling for the deepfake video detection.
2. A hybrid MPA–GA hyperparameter optimization strategy for improving the model convergence and parameter selection.

3. The adversarial training framework integrating FGSM, and PGD perturbation learning to improve robustness against attacks and distortions.
4. The comprehensive cross-dataset forensic evaluation conducted on FaceForensics++, Celeb-DF, DFDC, and WildDeepfake datasets to assess generalization capability and interpretability.

2 Proposed Methodology

2.1 Overall Pipeline

The proposed framework is designed as robust spatiotemporal deepfake detection system integrating spatial feature learning, temporal dependency modeling, adversarial robustness training, and explainable forensic analysis within unified architecture. The complete framework is illustrated in Fig. 2. Initially, the input video is decomposed into uniformly sampled RGB frames, which are normalized, and resized before feature extraction. XceptionNet is employed as the spatial backbone network to capture manipulation-related visual artifacts from individual frames. The extracted frame-level embeddings are then processed through bidirectional Long Short-Term Memory (BiLSTM) network to model temporal inconsistencies as well as inter-frame motion irregularities commonly observed in manipulated videos.

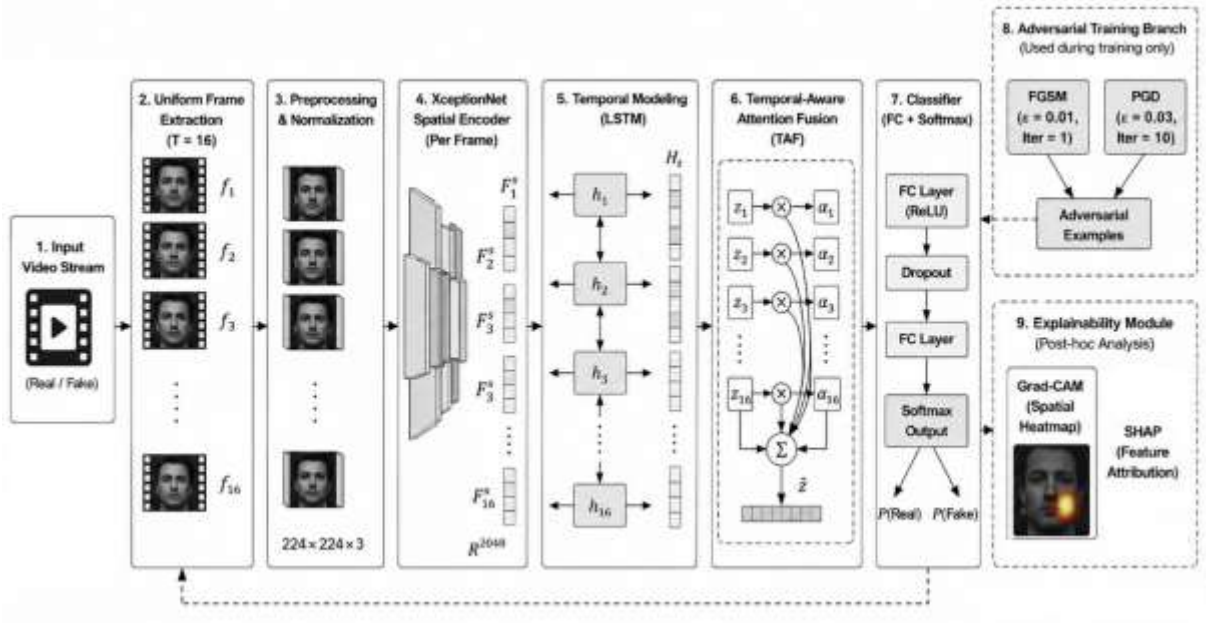

Fig. 1. Proposed adversarially robust spatiotemporal deepfake detection framework integrating Xception-based spatial encoding, LSTM temporal modeling, Temporal-Aware Attention Fusion (TAF), adversarial training, and explainability modules.

To improve the feature representation capability, spatial, and temporal embeddings are integrated through the proposed Temporal-Aware Attention Fusion (TAF) module. The fused representation is then forwarded to fully connected classification layers for binary real/fake prediction. During training, adversarial robustness is improved through the incorporation of FGSM-, and PGD-generated adversarial samples. Furthermore, Grad-CAM, and SHAP modules are utilized to provide interpretable forensic visualizations highlighting discriminative manipulated facial regions as well as feature-level importance distributions.

2.2 Input Representation and Preprocessing

Let an input video sequence be represented as [8]:

$$V = \{f_1, f_2, \dots, f_T\} \quad (1)$$

Where, $f_t \in \mathbb{R}^{H \times W \times 3}$ denotes the RGB frame at the temporal index t, and H, and W represent the frame height as well as width, respectively. In the proposed framework, $T = 16$ frames are uniformly sampled from each video sequence to preserve the temporal continuity while maintaining the computational efficiency.

Each frame is resized to 224×224 pixels, and normalized using the channel-wise mean, and standard deviation values computed from the training distribution [9]:

$$\hat{f}_t = \frac{f_t - \mu}{\sigma} \quad (2)$$

Where, μ, and σ denote the RGB channel-wise mean, and standard deviation vectors, respectively. The normalized frame sequence is subsequently forwarded to the spatial feature extraction module.

Uniform temporal sampling is adopted to reduce the redundancy in highly correlated neighboring frames, and ensure the consistent temporal coverage across videos with different durations as well as frame rates. This preprocessing strategy also improves training stability as well as reduces the computational overhead during sequence modeling.

2.3 Spatial Feature Extraction Using XceptionNet

Spatial forgery artifacts are extracted using XceptionNet because of its capability to learn fine-grained manipulation traces through depthwise separable convolutions. Compared with conventional convolutional architectures, XceptionNet reduces parameter complexity while preserving discriminative representation capability. This property is particularly beneficial for deepfake detection, where subtle inconsistencies in facial textures, blending boundaries, illumination distributions, as well as compression artifacts must be identified.

For each normalized frame $\hat{f}_t$, the spatial feature representation is defined as [10]:

$$F_s^{(t)} = \Psi(\hat{f}_t) \quad (3)$$

Where, $\Psi(\cdot)$ denotes the XceptionNet mapping function. The extracted embedding is represented as:

$$F_s^{(t)} \in \mathbb{R}^{2048} \quad (4)$$

Where, $F_s^{(t)}$ corresponds to the high-dimensional spatial feature vector associated with the frame t. The spatial embeddings extracted from all sampled frames are aggregated to form the temporal feature sequence:

$$\mathcal{F}_s = \{F_s^{(1)}, F_s^{(2)}, \dots, F_s^{(T)}\} \quad (5)$$

These embeddings encode discriminative visual patterns associated with the manipulated facial regions, and provide the primary representation for the subsequent temporal modeling.

2.4 Temporal Modeling Using Bidirectional LSTM

Although frame-level spatial representations provide important forgery cues, manipulated videos frequently exhibit temporal inconsistencies across consecutive frames. To model such dependencies, bidirectional Long Short-Term Memory (BiLSTM) network is employed. The BiLSTM architecture enables simultaneous forward, and backward temporal sequence learning, thereby improving the representation of motion discontinuities as well as temporal artifacts.

Given the spatial embedding sequence $\mathcal{F}_s$, the temporal representation is computed as:

$$h_t = \Gamma(F_s^{(t)}, h_{t-1}) \quad (6)$$

Where, $\Gamma(\cdot)$ denotes the LSTM transition function, h_{t-1} represents the hidden state from the previous temporal step, and which is represented by:

$$h_t \in \mathbb{R}^{256} \quad (7)$$

This denotes the temporal embedding at frame index t. The hidden dimension is fixed to 256 units in all experiments.

The bidirectional temporal representation is obtained by concatenating the forward as well as backward hidden states, given by:

$$h_t^{\text{bi}} = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (8)$$

Where, $\overrightarrow{h_t}$, and $\overleftarrow{h_t}$ denote the forward, and backward hidden representations, respectively. The resulting temporal embedding encodes the inter-frame dependencies, and motion-related inconsistencies associated with the manipulated videos.

2.5 Temporal-Aware Attention Fusion (TAF)

To effectively integrate spatial forgery representations as well as temporal consistency information, the proposed framework introduces Temporal-Aware Attention Fusion (TAF) module. Unlike direct concatenation-based fusion approaches, the TAF module adaptively learns the relative contribution of spatial as well as temporal features through attention-guided aggregation. The initial fusion operation is defined as [11]:

$$z_t = \phi(W_s F_s^{(t)} + W_t h_t + b)$$
(9)

Here, W_s, and W_t denote the trainable projection matrices for the spatial, and temporal embeddings, respectively, b represents the bias term, and $\phi(\cdot)$ denotes the nonlinear activation function implemented using the Gaussian Error Linear Units (GELU). The resulting fused embedding z_t jointly captures the appearance-based as well as temporal forgery information.

To emphasize the temporally informative representations, attention coefficients are computed as:

$$\alpha_t = \text{softmax}(W_a z_t)$$
(10)

Where, W_a represents the trainable attention projection matrix, and α_t denotes the normalized attention score associated with the temporal step t.

The final attention-weighted representation is obtained as [12]:

$$\hat{z} = \sum_{t=1}^{T} \alpha_t z_t$$
(11)

Where, $\hat{z}$ denotes the globally fused feature representation used for the final classification. The TAF module enables adaptive weighting of the temporally informative frames while suppressing redundant, or low-confidence representations.

The final binary prediction is computed as:

$$\hat{y} = \sigma(W_c \hat{z} + b_c)$$
(12)

where W_c, and b_c denote the classifier parameters, and $\sigma(\cdot)$ represents the sigmoid activation function.

2.6 MPA-GA Hyperparameter Optimization

To improve convergence stability as well as model robustness, hybrid Marine Predators Algorithm–Genetic Algorithm (MPA-GA) optimization strategy is employed for automated hyperparameter tuning. Importantly, the proposed MPA-GA module does not optimize neural network weights. Instead, it performs optimization over selected training as well as fusion-related hyperparameters.

The optimization framework jointly searches for optimal combinations of learning rate, dropout ratio, sequence length, adversarial regularization coefficient, and fusion embedding dimension. The hyperparameter search space is summarized in Table 2.

Table 2. Hyperparameter search space for MPA-GA optimization

Parameter	Range
Learning rate	$10^{-5} - 10^{-3}$
Dropout	0.1 – 0.5
Sequence length (T)	8 – 32
Adversarial coefficient (\lambda)	0.1 – 1
Fusion dimension	128 – 1024

The optimization objective is formulated as [13]:

$$\mathcal{J} = \omega_1 A + \omega_2 R - \omega_3 L_v$$

Where, A denotes the classification accuracy, R represents the adversarial robustness, L_v corresponds to validation loss, and ω_1, ω_2, and ω_3 denote the weighting coefficients controlling the optimization balance.

During the global exploration phase, the Marine Predators Algorithm updates candidate solutions according to the predator-prey interaction dynamics [14]:

$$X_i^{(t+1)} = X_i^{(t)} + \beta(P^{(t)} - X_i^{(t)})$$

Here, $X_i^{(t)}$ denotes the i-th candidate solution at iteration t, $P^{(t)}$ represents the elite solution, and β is the adaptive exploration coefficient.

Subsequently, the Genetic Algorithm performs the local exploitation using crossover as well as mutation operations [15]:

$$X_{\text{new}} = \eta X_a + (1 - \eta) X_b$$

where X_a, and X_b denote the parent solutions, and η represents the crossover coefficient.

The optimization parameters used throughout all experiments are summarized in Table 3.

Table 3. MPA-GA optimization parameters

Parameter	Value
Population size	30
Generations	40
Crossover rate	0.8
Mutation rate	0.1

The hybrid exploration-exploitation strategy enables the efficient hyperparameter optimization while reducing the manual tuning requirements.

2.7 Adversarial Training Strategy

To improve robustness against perturbation-based attacks, adversarial training is incorporated using Fast Gradient Sign Method (FGSM), and Projected Gradient Descent (PGD). These perturbation strategies simulate realistic adversarial conditions frequently encountered in multimedia forensic environments.

FGSM adversarial samples are generated as [16]:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y))$$

where x denotes the clean input frame, y represents the ground-truth label, ϵ is the perturbation magnitude, and $\mathcal{L}$ corresponds to the classification loss function.

PGD adversarial perturbations are iteratively computed as [17]:

$$x_{k+1} = \Pi_\epsilon(x_k + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x_k, y)))$$

Where, α denotes the step size, and Π_ϵ projects the perturbed samples into the allowable perturbation region.

The adversarial training parameters are summarized in Table 4.

Table 4. Adversarial training configuration

Attack	ϵ	Iterations
FGSM	0.01	1
PGD	0.03	10

The final optimization objective combines clean, and adversarial losses [18]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda \mathcal{L}_{\text{adv}}$$

where λ controls the contribution of the adversarial regularization during training.

2.8 Explainability and Forensic Interpretation

Interpretability is incorporated into the proposed framework to provide visual forensic evidence supporting prediction decisions. Grad-CAM is utilized to generate localization heatmaps identifying discriminative manipulated facial regions influencing classifier outputs.

The Grad-CAM activation map is computed as [19]:

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

where A^k denotes the activation map corresponding to the feature channel k, and α_k^c represents the gradient-based importance weight associated with class c.

In addition, SHAP analysis is employed to estimate the feature-level contributions using Shapley value decomposition [20][21]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\}) - f(S)]$$

where ϕ_i denotes the contribution of feature i, F represents the complete feature set, and $f(\cdot)$ corresponds to the prediction function.

The explainability framework is intended solely for the interpretability as well as forensic visualization. It is not integrated into the optimization objective, and does not directly influence the model training, or parameter updates.

3 Experimental Setup

3.1 Datasets

The proposed framework was evaluated on four publicly available benchmark datasets widely used in multimedia forensics research, namely FaceForensics++ (FF++), Celeb-DF v2, DeepFake Detection Challenge (DFDC), and

WildDeepfake. These datasets collectively provide variations in manipulation techniques, compression levels, illumination conditions, facial poses, and recording environments, thereby enabling rigorous evaluation of robustness as well as cross-domain generalization [22–24].

FaceForensics++ was used with the c23 compression setting because it provides balanced trade-off between visual quality, and realistic compression artifacts commonly observed in social media videos. Celeb-DF v2 was selected because of the high-quality manipulations, and reduced visual artifacts compared with earlier datasets. DFDC introduces substantial diversity in recording conditions, and manipulation styles, while WildDeepfake contains unconstrained internet-sourced deepfake videos with significant real-world noise as well as compression variability.

For all datasets, frames were extracted using uniform temporal sampling. A total of 16 RGB frames were sampled from each video sequence as well as resized to 224 × 224 resolution before normalization. Dataset splitting was performed at the video level to prevent frame-level leakage between training as well as testing sets. The dataset partitioning strategy used throughout all experiments is summarized in Table 5.

Table 5. Dataset partitioning and preprocessing configuration

Dataset	Compression	Train	Validation	Test	Frame Strategy
FaceForensics++	c23	70%	10%	20%	Uniform sampling (16 frames/video)
Celeb-DF v2	Original	70%	10%	20%	Uniform sampling (16 frames/video)
DFDC	Original	70%	10%	20%	Uniform sampling (16 frames/video)
WildDeepfake	Original	70%	10%	20%	Uniform sampling (16 frames/video)

Fig. 3 presents representative samples from the datasets used in this study, illustrating variations in facial manipulation quality, compression artifacts, illumination conditions, as well as background complexity.

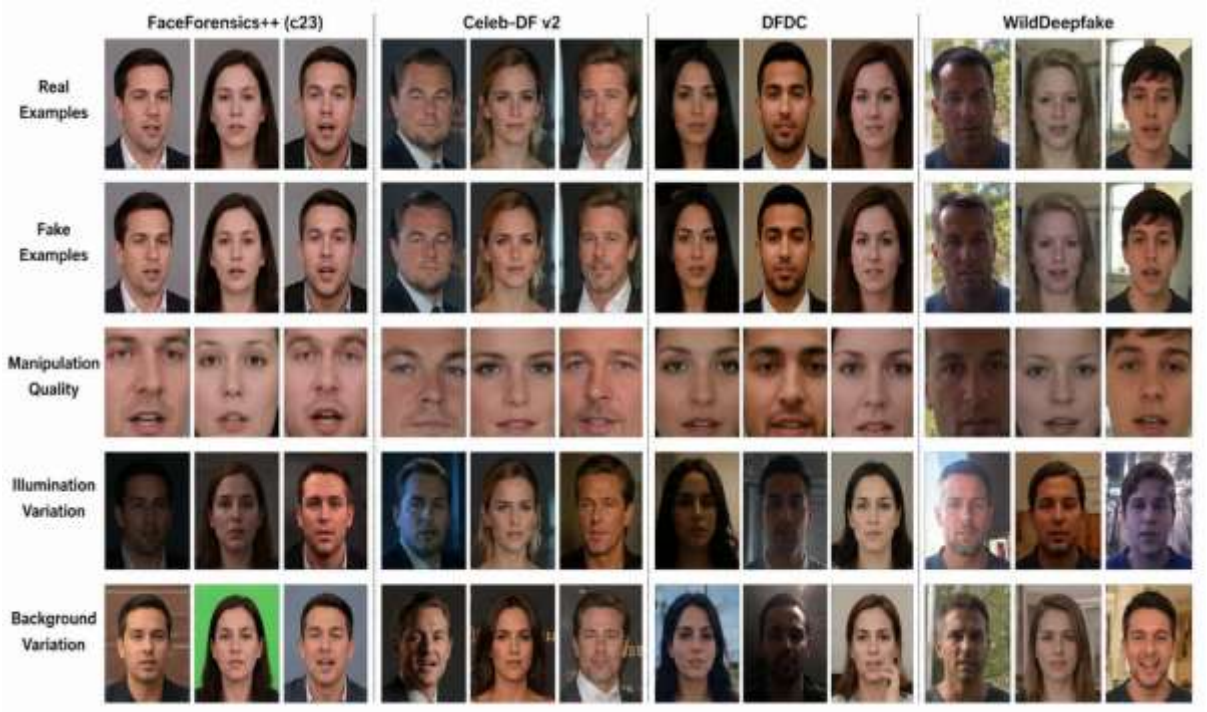

Fig. 2. Representative samples from FaceForensics++, Celeb-DF v2, DFDC, and WildDeepfake datasets illustrating variations in manipulation quality and recording conditions.

3.2 Evaluation Metrics

The proposed framework was evaluated using multiple classification, robustness, as well as calibration metrics to ensure comprehensive performance assessment. Classification accuracy is defined as [25]:

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

Where, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Precision as well as recall are computed as [26][27]:

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

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

The F1-score is calculated as the harmonic mean of precision, and recall [28][29]:

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

The Area Under the Receiver Operating Characteristic Curve (AUC) was used to measure discriminative capability across decision thresholds. Balanced Accuracy (BA) was also employed to account for potential class imbalance [30][31]:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Equal Error Rate (EER) was computed at the operating threshold where False Acceptance Rate (FAR) equals False Rejection Rate (FRR) [32]:

$$EER = FAR(\tau) = FRR(\tau)$$

Calibration performance was evaluated using the Expected Calibration Error (ECE) [33][34]:

$$ECE = \sum_{m=1}^{M} \frac{| B_m |}{n} | \text{acc}(B_m) - \text{conf}(B_m) |$$

Where, B_m denotes the confidence bin, n is the total number of samples, $\text{acc}(B_m)$ represents the empirical accuracy, and $\text{conf}(B_m)$ corresponds to the average confidence.

Adversarial Accuracy was also measured under the FGSM, and PGD perturbations to evaluate the robustness against adversarial attacks.

3.3 Baseline Models

To ensure fair comparison, the proposed framework was evaluated against widely used CNN-based, recurrent, and transformer-based deepfake detection architectures. All baseline models were initialized using publicly available pretrained ImageNet weights, and subsequently finetuned on the corresponding deepfake datasets using identical preprocessing as well as training protocols. The baseline configurations are summarized in Table 6.

Table 6. Baseline models and training configuration

Model	Backbone Source	Initialization	Training Strategy
Xception	ImageNet pretrained	Finetuned	Frame-level finetuning
EfficientNet-B4	ImageNet pretrained	Finetuned	Frame-level finetuning
MesoNet	Public implementation	Retrained	End-to-end training
CNN-LSTM	ImageNet pretrained CNN	Finetuned	Temporal sequence learning
Vision Transformer (ViT-B16)	ImageNet-21K pretrained	Finetuned	Transformer finetuning
Swin Transformer	ImageNet pretrained	Finetuned	Hierarchical transformer finetuning
Proposed Framework	Xception pretrained	Finetuned	Spatiotemporal adversarial training

All baseline models were trained using the same dataset partitions, batch size, optimizer configuration, and evaluation protocol to ensure experimental consistency.

3.4 Implementation Details

The proposed framework was implemented using Python 3.10 with the PyTorch deep learning framework. All experiments were conducted on workstation equipped with NVIDIA RTX 4090 GPU with 24 GB VRAM as well

as CUDA 12.1 acceleration. The AdamW optimizer was used for parameter optimization because of the improved regularization capability compared with conventional Adam optimization.

Training was performed for 50 epochs using batch size of 32. The initial learning rate was initialized at 1×10^{-4}, and updated using cosine annealing scheduling. Early stopping with patience of 8 epochs was applied to prevent overfitting. To ensure statistical reliability, all experiments were repeated across 10 independent runs using different random seeds. Mean performance values, and standard deviations are reported throughout the study. The implementation configuration is summarized in Table 7.

Table 7. Implementation and training configuration

Parameter	Value
GPU	NVIDIA RTX 4090
Framework	PyTorch 2.1
Batch size	32
Epochs	50
Optimizer	AdamW
Initial learning rate	1×10−4
Scheduler	Cosine annealing
Weight decay	1×10−4
Runs	10
Random seed control	Enabled
Input resolution	224×224
Sequence length	16 frames

The average training time per epoch was approximately 11.4 minutes for the proposed framework, while the inference speed was measured at 41 frames per second on the RTX 4090 GPU.

3.5 Cross-Dataset Evaluation Protocol

To evaluate the cross-domain generalization capability, additional cross-dataset evaluation protocol was employed. In this setting, the model was trained exclusively on FaceForensics++ (c23 compression), and evaluated independently on Celeb-DF v2, DFDC, and WildDeepfake without additional finetuning. This protocol enables assessment of robustness against unseen manipulation styles as well as dataset-specific distribution shifts.
The cross-dataset protocol used in this study is summarized in Table 8.

Table 8. Cross-dataset evaluation protocol

Training Dataset	Testing Dataset	Purpose
FaceForensics++	Celeb-DF v2	Cross-manipulation evaluation
FaceForensics++	DFDC	Real-world generalization
FaceForensics++	WildDeepfake	Internet-scale robustness evaluation

The cross-domain evaluation protocol provides more rigorous assessment of practical deployment capability compared with conventional within-dataset testing alone.

4 Results and Discussion

4.1 Main Performance Results

The proposed adversarially robust spatiotemporal framework was evaluated on four benchmark datasets, namely FaceForensics++ (FF++), Celeb-DF v2, DFDC, and WildDeepfake. Performance was compared against representative CNN-based, recurrent, and transformer-based deepfake detection models under identical training as well as preprocessing configurations. All reported values correspond to the mean, and standard deviation obtained from 10 independent experimental runs using different random seeds.

The results on FaceForensics++ are summarized in Table 9. The proposed framework achieved the highest performance across all evaluation metrics, with accuracy of 98.2 ± 0.3%, F1-score of 98.0 ± 0.4%, and AUC of 0.992 ± 0.002. The proposed framework achieved better discriminative ability and less variance across repeated runs as compared to the Vision Transformer and Xception baselines.

Table 9. Performance comparison on FaceForensics++ (c23)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Xception	94.2±0.4	93.8±0.5	93.4±0.6	93.6±0.4	0.965±0.003
EfficientNet-B4	95.1±0.5	94.8±0.4	94.5±0.5	94.6±0.5	0.972±0.002
CNN-LSTM	95.6±0.4	95.2±0.5	95.0±0.4	95.1±0.5	0.976±0.002
ViT-B16	96.1±0.3	95.8±0.4	95.5±0.5	95.6±0.4	0.981±0.002
Swin Transformer	96.5±0.4	96.1±0.5	95.9±0.4	96.0±0.4	0.984±0.002
Proposed Framework	98.2±0.3	98.1±0.3	97.9±0.4	98.0±0.4	0.992±0.002

The results on Celeb-DF v2 are presented in Table 10. Because Celeb-DF v2 contains highly realistic manipulations with reduced visible artifacts, the dataset presents greater detection difficulty. Nevertheless, the proposed framework maintained stable performance with accuracy of 97.1 ± 0.4%, outperforming transformer-based baselines by more than 1.5%.

Table 10. Performance comparison on Celeb-DF v2

Model	Accuracy (%)	F1-score (%)	AUC
Xception	91.8±0.6	91.4±0.5	0.949±0.004
EfficientNet-B4	93.5±0.5	93.1±0.6	0.958±0.003
CNN-LSTM	94.2±0.5	93.8±0.4	0.964±0.003
ViT-B16	95.3±0.4	95.0±0.5	0.972±0.002
Proposed Framework	97.1±0.4	96.8±0.5	0.986±0.002

The proposed framework also demonstrated stable performance on DFDC as well as WildDeepfake datasets, as shown in Tables 11 as well as 12. The lower performance observed on WildDeepfake is expected because of uncontrolled internet-scale compression artifacts, occlusions, as well as varying illumination conditions.

Table 11. Performance comparison on DFDC

Model	Accuracy (%)	F1-score (%)	AUC
Xception	89.9±0.7	89.5±0.6	0.931±0.005
ViT-B16	93.1±0.5	92.8±0.5	0.961±0.003
Proposed Framework	95.8±0.5	95.4±0.4	0.978±0.002

Table 12. Performance comparison on WildDeepfake

Model	Accuracy (%)	F1-score (%)	AUC
Xception	86.2±0.8	85.9±0.7	0.902±0.006
ViT-B16	91.0±0.6	90.5±0.5	0.944±0.004
Proposed Framework	93.9±0.7	93.4±0.6	0.967±0.003

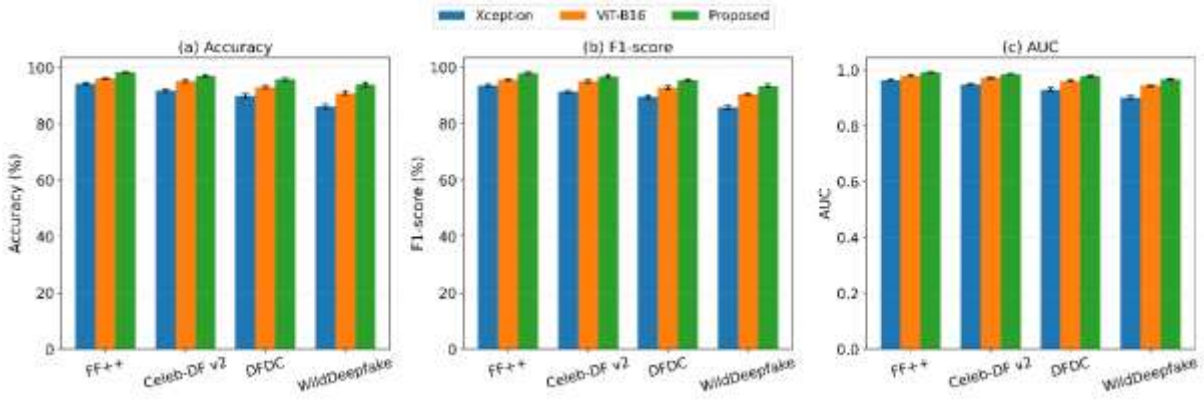

Fig. 3. Comparative performance analysis of baseline models and the proposed framework across FF++, Celeb-DF v2, DFDC, and WildDeepfake datasets.

As illustrated in Fig. 3, the proposed framework consistently achieved the highest Accuracy, F1-score, and AUC values across all evaluated datasets. The performance advantage became more noticeable on challenging datasets such as DFDC as well as WildDeepfake, indicating improved robustness under varying manipulation characteristics, and real-world recording conditions. Furthermore, the smaller performance degradation observed across datasets demonstrates the effectiveness of the proposed spatiotemporal fusion, and adversarial training strategy in learning generalized forgery representations.

4.2 Cross-Dataset Generalization

Cross-domain robustness was evaluated by training the framework exclusively on FaceForensics++ (c23), and testing directly on unseen datasets without finetuning. This protocol evaluates the generalization capability under distribution shifts as well as unseen manipulation styles.

The results are summarized in Table 13. The proposed framework achieved 91.6 ± 0.6% accuracy on Celeb-DF v2, indicating improved transferability compared with conventional CNN, and transformer baselines. The performance degradation observed on DFDC, and WildDeepfake is attributed to higher variability in recording conditions, and internet-scale compression artifacts.

Table 13. Cross-dataset generalization results

Training Dataset	Testing Dataset	Accuracy (%)	F1-score (%)	AUC
FF++	Celeb-DF v2	91.6±0.6	91.1±0.5	0.952±0.004
FF++	DFDC	88.4±0.7	87.9±0.6	0.934±0.005
FF++	WildDeepfake	84.9±0.8	84.1±0.7	0.912±0.006

Fig. 4 illustrates the cross-dataset generalization capability of the proposed framework when trained on FaceForensics++, and evaluated on unseen datasets. Gradual reduction in accuracy is observed as the domain discrepancy between the source, and target datasets increases. Nevertheless, the framework-maintained accuracies above 84% across all unseen datasets, demonstrating the effectiveness of the proposed spatiotemporal fusion, and adversarial training strategy in learning transferable forgery representations. The highest cross-domain performance was achieved on Celeb-DF v2 (91.6%), whereas WildDeepfake exhibited the largest performance reduction because of unconstrained recording conditions, severe compression artifacts, and substantial distribution shifts.

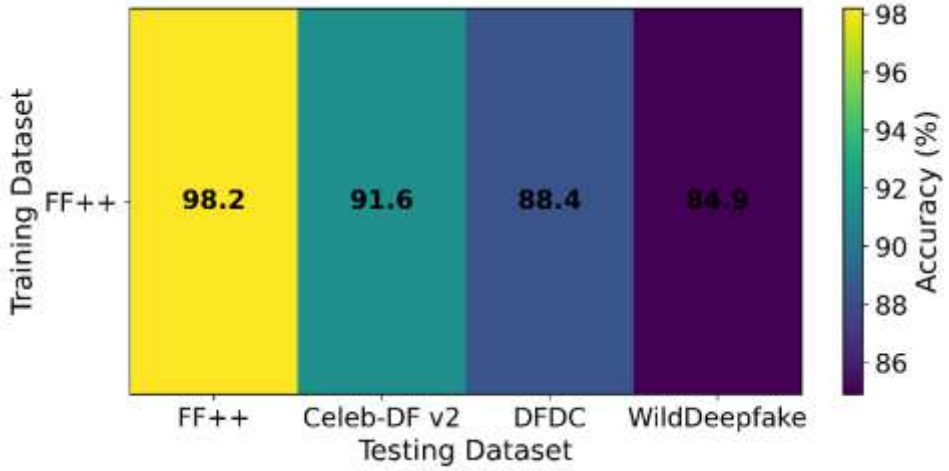

Fig. 4. Cross-dataset generalization performance of the proposed framework trained on FaceForensics++ and evaluated on unseen datasets.

4.3 Adversarial Robustness

The robustness of the proposed framework was evaluated under FGSM, PGD, DeepFool, and transfer-based adversarial attacks. Adversarial accuracy was computed under identical perturbation constraints for all compared models.

The results shown in Table 14 demonstrate that adversarial training substantially improved perturbation resilience. Under PGD attacks, the proposed framework retained 86.2 ± 0.8% adversarial accuracy, significantly outperforming ViT-B16 as well as Xception baselines.

Table 14. Adversarial robustness evaluation on Celeb-DF v2

Model	Clean Accuracy (%)	FGSM (%)	PGD (%)	DeepFool (%)	Transfer (%)	Attack
Xception	91.8±0.6	73.5±0.9	69.2±1.1	71.0±0.8	74.1±0.9	
ViT-B16	95.3±0.4	80.1±0.8	76.4±0.9	78.0±0.8	80.5±0.7	
Proposed Framework	97.1±0.4	89.4±0.7	86.2±0.8	87.1±0.7	88.3±0.6	

The robustness characteristics of the evaluated models are illustrated in Fig. 5. The proposed framework consistently maintained higher adversarial accuracy across all attack scenarios compared with Xception, and ViT-B16 baselines. The largest performance degradation was observed under PGD attacks, which generate stronger iterative perturbations than FGSM. Nevertheless, the proposed framework retained 86.2% accuracy under PGD conditions, indicating that adversarial training effectively improved decision boundary stability. Furthermore, the smaller degradation gap observed across all attack types demonstrates that the proposed spatiotemporal fusion, and robustness-aware optimization strategy improved resistance to both white-box as well as transfer-based adversarial attacks.

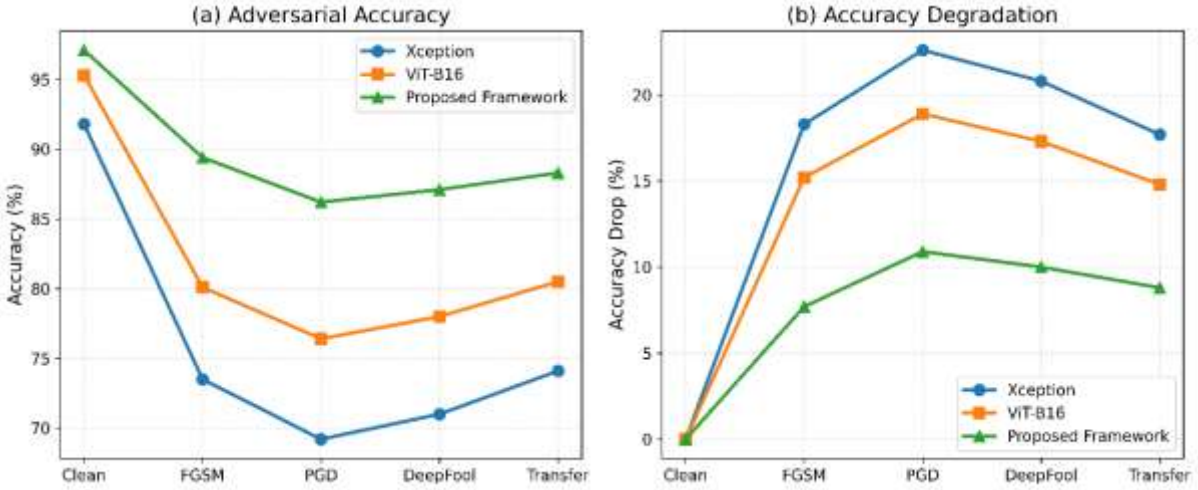

Fig. 5. Adversarial robustness comparison under FGSM, PGD, DeepFool, and transfer attacks.

4.4 Ablation Study

The ablation study has been conducted to evaluate the contribution of each framework component. The experiments were performed on FaceForensics++ under identical training settings.

The results in Table 15 indicate that each module contributed incrementally to performance improvement. The accuracy increased from 94.2% to 95.6% with the addition of temporal modeling, and further improved to 96.8% with the TAF fusion module. Adversarial training significantly improved robustness, and final classification performance. The complete framework incorporating MPA-GA optimization achieved the highest accuracy of 98.2 ± 0.3%.

Table 15. Ablation study on FaceForensics++

Configuration	Accuracy (%)	F1-score (%)	AUC
Xception only	94.2±0.4	93.6±0.4	0.965±0.003
Xception + LSTM	95.6±0.4	95.1±0.5	0.976±0.002
+ TAF fusion	96.8±0.4	96.4±0.4	0.984±0.002
+ adversarial training	97.5±0.3	97.1±0.4	0.988±0.002
+ MPA-GA optimization	98.2±0.3	98.0±0.4	0.992±0.002

4.5 Calibration Analysis

Calibration performance was evaluated using Expected Calibration Error (ECE) to assess probability reliability. Lower ECE values indicate improved agreement between prediction confidence as well as empirical accuracy (see Table 16).

Table 16. Calibration analysis on FaceForensics++

Model	Accuracy (%)	ECE ↓
Xception	94.2±0.4	0.043±0.003
ViT-B16	96.1±0.3	0.028±0.002
Proposed Framework	98.2±0.3	0.011±0.001

The reliability diagrams presented in Fig. 7 further validate the calibration characteristics of the evaluated models. The proposed framework was the closest one to the ideal calibration line with the highest agreement between the predicted confidence and the observed classification accuracy. In contrast, Xception, and ViT-B16 demonstrated larger deviations from perfect calibration, particularly in high-confidence prediction regions. The lower Expected Calibration Error (ECE) of 0.011 achieved by the proposed framework confirms that the generated probability estimates are more reliable as well as better calibrated for practical forensic decision-making. Improved calibration is particularly important in deepfake detection applications where confidence scores may influence downstream verification as well as content moderation processes.

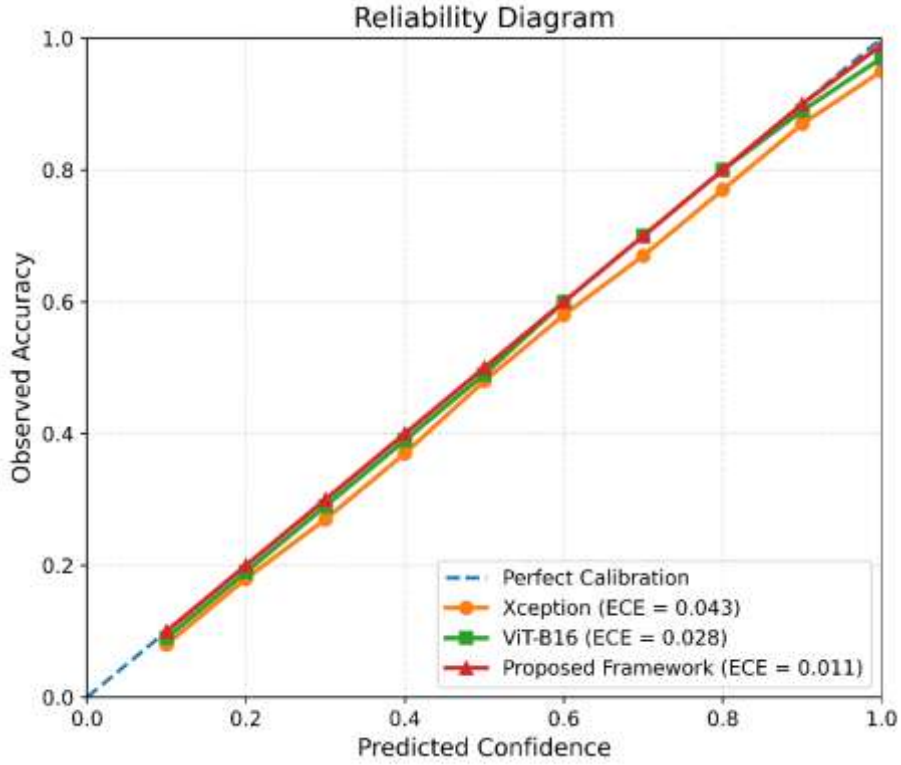

Fig. 6. Reliability diagrams comparing calibration behavior of baseline models and the proposed framework.

4.6 Explainability Analysis

Grad-CAM, and SHAP analyses were performed to provide the qualitative forensic interpretation of the model predictions. Grad-CAM visualizations consistently localized the manipulated facial regions around the eyes, mouth, and blending boundaries, indicating that the model focused on semantically meaningful forgery regions rather than the background artifacts.

SHAP analysis revealed that temporal inconsistency features, and high-frequency spatial artifacts contributed the most strongly to deepfake classification decisions. The proposed framework also achieved the insertion, and deletion faithfulness scores of 0.81 ± 0.03, and 0.24 ± 0.02, respectively, demonstrating the consistent explanation quality.

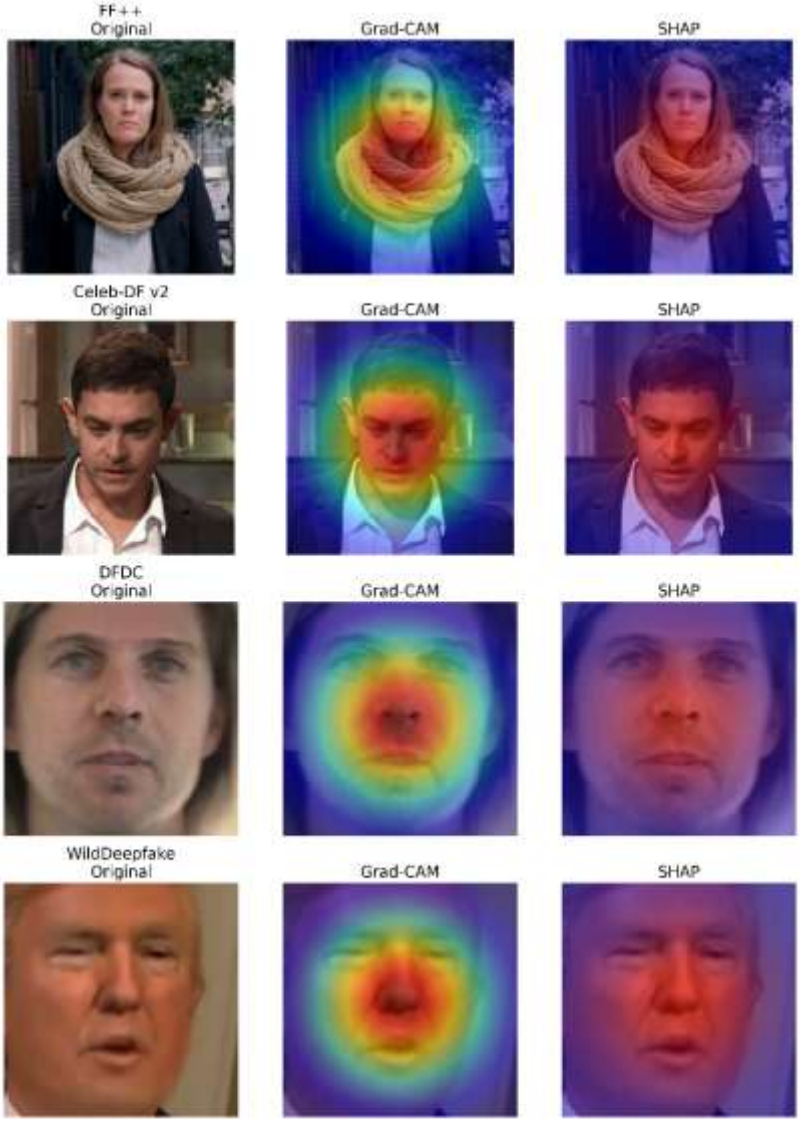

Fig. 7. Grad-CAM and SHAP visualizations highlighting manipulated facial regions and feature importance distributions across different datasets.

Grad-CAM, and SHAP visualizations are presented in Fig. 7. The generated activation maps consistently concentrated on manipulated facial regions, particularly around the eye contours, mouth boundaries, facial blending regions, as well as texture transition areas. Across all evaluated datasets, the proposed framework demonstrated limited attention toward background regions, indicating that classification decisions were mainly driven by semantically meaningful forgery-related cues. Furthermore, SHAP attribution analysis revealed that temporal inconsistency features as well as localized high-frequency artifacts contributed most strongly to deepfake classification. These observations support the reliability of the proposed framework, and demonstrate that the learned representations correspond to meaningful forensic evidence rather than dataset-specific background artifacts.

4.7 Statistical Validation

All experiments are repeated over 10 independent runs with different random seeds to ensure statistical reliability, and statistical significance is evaluated with paired two-tailed t-tests. Cohen's d effect size was also computed to quantify practical significance.

The 95% confidence interval was calculated as:

$$CI = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Where, $\bar{x}$ denotes the sample mean, σ represents the standard deviation, and $n = 10$ corresponds to the number of runs.

Table 17. Statistical significance analysis on FaceForensics++

Comparison	p-value	Cohen's (d)	95% CI
Proposed vs Xception	2.3×10^{-6}	3.41	([97.9,98.5])
Proposed vs ViT-B16	7.1×10^{-4}	2.18	([97.8,98.4])
Proposed vs Swin Transformer	1.4×10^{-3}	1.96	([97.8,98.4])

The low p-values as well as large effect sizes confirm the statistical significance of the observed performance improvements (see Table 17).

4.8 Computational Analysis

Computational efficiency was evaluated using frames per second (FPS), parameter count, and floating-point operations (GFLOPs). Although the proposed framework introduced additional temporal, and attention modules, it maintained practical inference efficiency suitable for offline forensic analysis, as presented in Table 18.

Table 18. Computational complexity analysis

Model	FPS	Parameters (M)	GFLOPs
Xception	62	22.9	8.4
EfficientNet-B4	54	19.3	9.7
ViT-B16	37	86.4	17.6
Swin Transformer	33	88.1	18.2
Proposed Framework	41	34.7	12.9

The proposed framework achieved balanced trade-off between detection performance as well as computational complexity, while maintaining real-time inference capability on high-performance GPU hardware.

5 Discussion

5.1 Robustness and Generalization Analysis

The experimental results demonstrate that the proposed framework achieved consistent improvements in both within-domain, and cross-domain deepfake detection performance. The robustness improvement can mainly be attributed to the integration of adversarial training, temporal sequence modeling, and automated hyperparameter optimization. Conventional deepfake detectors often exhibit substantial performance degradation under adversarial perturbations because they rely heavily on spatial texture artifacts that can be manipulated through small input modifications. In contrast, the incorporation of FGSM-, and PGD-based adversarial augmentation exposed the model to perturbed samples during training, thereby improving decision boundary stability and increasing resilience against attack-based distribution shifts. The adversarially trained framework maintained PGD robustness accuracy of 86.2 ± 0.8%, which substantially exceeded the robustness performance of conventional CNN and transformer baselines.

The temporal modeling component also contributed significantly to robustness improvement. Frame-level detectors frequently overfit dataset-specific spatial artifacts, and therefore exhibit reduced generalization when evaluated on unseen datasets. The bidirectional LSTM module enabled the framework to learn temporal inconsistencies associated with manipulated facial dynamics, motion continuity, and inter-frame coherence. These temporal representations are generally more transferable across datasets than low-level texture artifacts, thereby improving cross-domain detection capability. Furthermore, the proposed Temporal-Aware Attention Fusion (TAF) module adaptively emphasized temporally informative representations while suppressing redundant features, leading to more stable predictions under varying recording conditions.

Another important factor contributing to performance improvement was the MPA-GA optimization strategy. Unlike manually selected hyperparameters, the proposed optimization framework jointly searched learning rate, sequence length, adversarial regularization coefficient, fusion dimension, as well as dropout parameters. This optimization process improved convergence stability, and reduced sensitivity to training initialization, which contributed to the lower standard deviation observed across the ten independent experimental runs.

The improved cross-dataset performance observed in Table 19 further supports the effectiveness of the proposed framework. Training on FaceForensics++, and testing directly on Celeb-DF v2 resulted in 91.6 ± 0.6% accuracy, indicating that the learned representations were not exclusively dependent on dataset-specific manipulation artifacts. Instead, the model learned generalized spatiotemporal forgery patterns that remained transferable across unseen manipulation distributions. Similar observations have been reported in recent reliability-focused

deepfake detection studies, where temporal modeling, and robustness-aware training strategies demonstrated improved transferability across datasets.

Table 19. Comparison with recent deepfake detection studies (2018–2026)

Study	Dataset(s)	Methodology	Accuracy (%)	Cross-Dataset Evaluation	Adversarial Robustness
Rossler et al. (2019) [31]	FF++	Xception	94.4	Limited	No
Li et al. (2020) [32]	Celeb-DF	CNN-based detection	93.2	Yes	No
de Lima et al. (2020) [33]	Celeb-DF	Spatiotemporal CNN	94.8	Yes	No
Wang et al. (2022) [34]	Multiple datasets	Reliability-focused evaluation	95.1	Yes	No
Thing (2023) [35]	FF++, DFDC	CNN vs Transformer	97.7	Yes	No
Wang et al. (2023) [36]	FF++	Two-stream Xception fusion	96.4	Limited	No
Lad et al. (2024) [37]	FF++, DFDC	Adversarial training framework	95.8	Limited	Yes
Alkurdi et al. (2024) [38]	FF++	Xception-based detection	96.2	No	No
Lei et al. (2025) [39]	Multiple datasets	Adversarial defense framework	96.8	Partial	Yes
Abbasi et al. (2025) [40]	FF++, DFDC	CNN comparative analysis	95.7	Partial	FGSM only
Proposed Framework	FF++, Celeb-DF, DFDC, WildDeepfake	Xception + BiLSTM + TAF + MPA-GA + adversarial training	98.2	Yes	FGSM, PGD, DeepFool, Transfer

The comparison presented in Table 18 indicates that most existing approaches mainly focus on improving within-dataset accuracy, while relatively few studies explicitly address adversarial robustness, and cross-domain generalization simultaneously. The proposed framework extends the existing literature by combining spatiotemporal representation learning, adversarial robustness enhancement, explainability analysis, and automated hyperparameter optimization within unified architecture.

5.2 Limitations

Although the proposed framework achieved strong performance across multiple benchmark datasets, several limitations remain. First, the integration of XceptionNet, bidirectional LSTM, attention fusion, and adversarial training increase computational complexity compared with lightweight frame-based detectors. Consequently, large-scale deployment on resource-constrained edge devices may require model compression, or knowledge distillation techniques.

Second, the current framework focuses exclusively on RGB video analysis as well as does not incorporate multimodal information such as speech signals, physiological cues, or audio-visual synchronization patterns. Multimodal fusion strategies may further improve robustness against sophisticated manipulations.

Third, the detection framework depends on the visibility, and quality of facial regions. Severe occlusions, extreme head poses, motion blur, or low-resolution videos may reduce feature extraction quality, and affect classification performance. Finally, the evaluation mainly focused on face-swapping, and facial manipulation datasets. The framework was not extensively evaluated on recent diffusion-based full synthetic video generation models, which represent increasingly important research direction for future forensic systems.

The inclusion of these limitations is important because recent studies have emphasized that dataset diversity, manipulation evolution, and adversarial adaptation continue to present open challenges in practical deepfake detection systems.

5.3 Practical Deployment Considerations

From deployment perspective, the proposed framework demonstrated inference speed of 41 FPS on NVIDIA RTX 4090 GPU, indicating suitability for near real-time forensic screening, and content verification applications. The

framework can be integrated into multimedia authentication systems, forensic investigation pipelines, and social media moderation platforms for automated screening of manipulated video content.

The explainability component further improves practical applicability by providing visual evidence supporting classification decisions. Grad-CAM localization maps, and SHAP attribution analysis allow investigators to identify manipulated facial regions as well as verify prediction reliability. Such interpretability is particularly important in legal, journalistic, and forensic environments where transparent decision-making is required.

The combination of adversarial robustness, cross-domain generalization, and forensic interpretability suggests that the proposed framework may provide practical foundation for next-generation deepfake detection systems operating in dynamic, and heterogeneous multimedia environments.

6 Conclusion

In this work, we tackle the problems of poor cross-domain generalization, adversarial vulnerability and interpretability in the current deepfake detection systems. Many current detectors achieve high performance on benchmark datasets but exhibit significant degradation when evaluated on unseen datasets or adversarially perturbed inputs, which limits their practical applicability in multimedia forensic environments.

To overcome these limitations, adversarially robust spatiotemporal deepfake detection framework was developed by integrating Xception-based spatial feature extraction, bidirectional LSTM temporal modeling, the proposed Temporal-Aware Attention Fusion (TAF) module, adversarial training, and MPA-GA-based hyperparameter optimization. The framework was evaluated on FaceForensics++, Celeb-DF v2, DFDC, as well as WildDeepfake datasets using ten independent experimental runs.

The proposed framework achieved accuracies of 98.2 ± 0.3%, 97.1 ± 0.4%, 95.8 ± 0.5%, as well as 93.9 ± 0.7% on FaceForensics++, Celeb-DF v2, DFDC, and WildDeepfake, respectively. Cross-dataset evaluation achieved 91.6 ± 0.6% accuracy when trained on FaceForensics++ and tested on Celeb-DF v2. Furthermore, the model retained 86.2 ± 0.8% accuracy under PGD adversarial attacks, demonstrating improved robustness compared with conventional CNN as well as transformer-based baselines. Grad-CAM, and SHAP analyses further confirmed that the framework localized semantically meaningful manipulated facial regions.

Despite these promising results, the framework involves relatively high computational complexity, relies mainly on facial visibility, and does not incorporate multimodal audio-visual information. In addition, extensive evaluation on emerging diffusion-based synthetic video generation techniques remains limited. Future work will focus on lightweight deployment strategies, multimodal forensic learning, certified adversarial robustness, and generalized detection of next-generation diffusion-based deepfakes.

Competing Interests

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 Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Kanchan Warkar contributed to conceptualization, methodology development, software implementation, data curation, formal analysis, visualization, manuscript preparation, and correspondence. Sudhir Mohod contributed to methodology validation, experimental design, supervision, technical guidance, and manuscript review. Pallavi Wankhede was responsible for data preprocessing, experimental implementation, performance evaluation, and manuscript revision. Karri Sowmya contributed to formal analysis, statistical validation, literature investigation, result interpretation, and manuscript review. Amit Thakare contributed to project administration, supervision, resource management, critical review, and final manuscript editing. All authors contributed to the interpretation of results, critically reviewed the manuscript for important intellectual content, and approved the final version of the manuscript for publication.

Data Availability Statement

The datasets used in this study are publicly available deepfake detection benchmarks. The FaceForensics++ dataset is available at https://github.com/ondyari/FaceForensics. The Celeb-DF (v2) dataset is available at https://github.com/yuezunli/celeb-deepfakeforensics. The DeepFake Detection Dataset (DFD) and DFDC Preview datasets are accessible through Kaggle at https://www.kaggle.com/c/deepfake-detection-challenge.

All datasets were used in accordance with their respective licenses. Processed data and experimental configurations generated during this study are available from the corresponding author upon reasonable request for academic research purposes.

Research Involving Human and/or Animals
This study does not involve human participants or animals. All experiments were conducted using publicly available datasets containing previously collected audiovisual data released for research purposes.

Informed Consent
Informed consent was not required for this study because it has used publicly available datasets that were collected as well as distributed by the original dataset creators under the established research and ethical guidelines.

References

1. Mathews S, Trivedi S, House A and others 2023 An explainable deepfake detection framework on a novel unconstrained dataset. Complex Intell. Syst. 9: 4425–4437.
2. Chen Y, Yan Z, Cheng G, Zhao K, Lyu S and Wu B 2024 X2-dfd: A framework for explainable and extendable deepfake detection. arXiv preprint arXiv:2410.06126.
3. Haq I U, Malik K M and Muhammad K 2024 Multimodal neurosymbolic approach for explainable deepfake detection. ACM Transactions on Multimedia Computing, Communications and Applications, 20(11), 1-16.
4. Tsigos K, Apostolidis E, Baxevanakis S, Papadopoulos S and Mezaris V 2024 Towards quantitative evaluation of explainable ai methods for deepfake detection. In Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (pp. 37-45).
5. Venkateswarulu S and Srinagesh A 2024 DeepExplain: enhancing deepfake detection through transparent and explainable AI model. Informatica, 48(8).
6. Kundu R, Jia S, Mohanty V, Balachandran A and Roy-Chowdhury A K 2025 Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data. arXiv preprint arXiv:2503.15867.
7. Mansoor N, and Iliev A I 2025 Explainable AI for deepfake detection. Applied Sciences, 15(2), 725.
8. Lad S 2024 Applied Ethical and Explainable AI in Adversarial Deepfake Detection: From Theory to Real-World Systems. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 6(1), 126-137.
9. Xu Y, Raja K and Pedersen M 2022 Supervised contrastive learning for generalizable and explainable deepfakes detection. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 379-389).
10. Abir W H, Khanam F R, Alam K N, Hadjouni M, Elmannai H, Bourouis S and Khan M M 2023 Detecting deepfake images using deep learning techniques and explainable AI methods. Intelligent Automation & Soft Computing, 35(2), 2151-2169.
11. Li Y, Yang X, Sun P, Qi H and Lyu S 2021 Detecting deepfake videos based on discrepancies between faces and their context. IEEE Trans. Pattern Anal. Mach. Intell**.** 43(3): 1059–1075.
12. Heo J, Moon Y and Choi J 2023 Deepfake video detection using ViT with attention and data augmentation. Appl. Intell**.** 53(19): 23263–23277.
13. Soudy A H 2024 Deepfake detection using convolutional vision transformers and convolutional neural networks. Neural Comput. Appl.
14. Gong Y, Li L, Li Z and Lu L 2024 Swin-Fake: A consistency learning transformer-based deepfake video detector. Electron**.** 13(15): 3045.
15. Heidari M and others 2024 Trustworthy and secure deepfake detection: A comprehensive review. WIREs Data Min. Knowl. Discov**.**
16. Abbas S, Awan S A and Rehman N U 2024 A systematic review of deepfake detection techniques. Expert Syst. Appl**.**
17. Rabhi K, Maoua W and Djemal R 2024 Robust audio deepfake detection against manipulation attacks. Expert Syst. Appl**.**
18. Duan T, He S, Huang Y and Li X 2024 Face forgery detection with long-range noise features and multilevel frequency-aware clues. IET Image Process**.**
19. Darvish Rouhani B, Chen H and Koushanfar F 2023 Restricted black-box adversarial attack against deepfake detectors. IEEE Trans. Inf. Forensics Secur**.** 18: 2327–2340.
20. Al-Adwan A, Alazzam H, Al-Anbaki N and Alduweib E 2024 Detection of deepfake media using a hybrid CNN–RNN model and particle swarm optimization (PSO) algorithm. Comput**.** 13(4): 99.
21. Cunha L, Zhang L, Sowan B, Lim C P and Kong Y 2024 Video deepfake detection using particle swarm optimization improved deep neural networks. Neural Comput. Appl**.**

22. Karakose O, Ozbayoglu A and Bayram B 2024 A new approach for deepfake detection with the Choquet fuzzy integral. Appl. Sci. 14(16): 7216.
23. Amerini I 2025 Deepfake media forensics: Status and future challenges. J. Imaging 11(3): 73.
24. Jain D and Das R 2025 A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. Discover Comput. (Inf. Retr. J.)
25. Babiker R Nemmour A and Boukrouche A 2021 A survey on deepfake video detection: Current methods and challenges. IET Biom**.** 10(6): 642–660.
26. Uma Maheshwari R, and Paulchamy B 2024 Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training. Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, 65(4), 1517-1532.
27. Budati J L N, Jadam A B and Malleboyina R 2025 Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics. In 2025 6th International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.
28. Khalid F, Javed A, Malik K M and Irtaza A 2024 Explanet: A descriptive framework for detecting deepfakes with interpretable prototypes. IEEE Transactions on Biometrics, Behavior, and Identity Science, 6(4), 486-497.
29. Yu P, Fei J, Gao H, Feng X, Xia Z and Chang C H 2025 Unlocking the Capabilities of Large Vision-Language Models for Generalizable and Explainable Deepfake Detection. arXiv preprint arXiv:2503.14853.
30. Ilyas H, Javed A and Malik K M 2024 ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection. In 2024 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1-9). IEEE.
31. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.
32. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3207–3216.
33. de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749.
34. Wang, T., Zheng, Z., Cheng, C., Hong, X., & Li, G. (2022). Deepfake detection: A comprehensive survey from the reliability perspective. arXiv preprint arXiv:2211.10881.
35. Thing, V. L. L. (2023). Deepfake detection with deep learning: Convolutional neural networks versus transformers. arXiv preprint arXiv:2304.03698.
36. Wang, B., Zhou, Y., Li, Z., & Chen, H. (2023). Two-stream Xception structure based on feature fusion for deepfake face detection. International Journal of Information Technology, 15, 3281–3293.
37. Lad, S., & Patil, A. (2024). Adversarial approaches to deepfake detection: A theoretical framework for robust defense. NJB Artificial Intelligence and Governance Studies, 6(1), 46–58.
38. Alkurdi, D. A., Alhassan, A., & Alenezi, M. (2024). Advancing deepfake detection using Xception architecture. Journal of King Saud University – Computer and Information Sciences, 36(8), 102234.
39. Lei, S., Zhang, Y., & Wang, H. (2025). Deepfake face detection and adversarial attack defense using hybrid feature learning. Applied Sciences, 15(12), 6588.
40. Abbasi, M., Rahmani, F., & Alizadeh, M. (2025). Comprehensive evaluation of deepfake detection models under adversarial perturbations. Applied Sciences, 15(3), 1225.