Deepfake Detection Using ViT-BiLSTM Based Face Analysis

Gopavarapu Sri Rama Krishna Vamsi[1], Garlapati Ganesh[2], Busi Vineeth Kumar[3], Guntur Jaswanth[4], Nagababu Pachhala[5]
[1,2,3,4]*Students, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, India*
[5]*Associate Professor, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Guntur, India*

***Abstract*— Deepfake videos have become increasingly realistic due to advancements in generative models, posing significant challenges for digital media authenticity and security. This paper presents a hybrid deep learning framework for deepfake video detection that integrates Vision Transformer (ViT)-based spatial feature extraction with Bidirectional Long Short-Term Memory (BiLSTM) temporal modelling. The proposed system processes videos through frame extraction, face detection using Retina Face with MTCNN as fallback, identity-consistent face tracking using a multi-criterion approach, and feature extraction using a pretrained ViT model. The extracted embeddings are analysed using a BiLSTM network to capture temporal inconsistencies across frames. The model is trained on Face Forensics++ and CelebDF datasets and evaluated using both internal and cross-dataset validation. Experimental results show that the proposed approach achieves high accuracy, with a balanced accuracy of 93.5% and ROC-AUC of 0.987 on the internal test set, demonstrating strong discriminative capability and low false-positive rates. However, performance on SDFVD dataset indicates a reduction in generalization due to domain shift. The proposed system provides an effective and reliable solution for deepfake detection, particularly in forensic applications where minimizing false positives is critical, and highlights the need for improved cross-dataset robustness in future research.**

***Index Terms*—Deepfake Detection, Vision Transformer (ViT), Bidirectional LSTM (BiLSTM), Temporal Analysis, Multimedia Forensics, Face Tracking.**

I. INTRODUCTION

The proliferation of sophisticated generative adversarial networks and autoencoders has made it increasingly simple to create photorealistic synthetic media, commonly known as deepfakes [1]. These manipulated videos, particularly those targeting human faces, present major societal risks ranging from defamation and political destabilization to fraud. Reliable deepfake detection is therefore a critical research area in computer vision and multimedia forensics [2], [3].

Traditional detection methods often struggle due to the high visual quality and temporal consistency of modern deepfakes, as well as confounding factors such as video compression and varied lighting. While frame-based CNN approaches can identify spatial artifacts, they often fail to capture subtle inconsistencies that manifest over time.

This work addresses these limitations by employing a video-based temporal analysis strategy focused on facial regions where deepfake artifacts are most prominent.

Contribution: This work presents a complete deepfake detection pipeline leveraging identity-aware face tracking, a pretrained Vision Transformer for discriminative spatial embeddings, and a BiLSTM for temporal analysis. It provides a transparent evaluation of cross-dataset generalization via SDFVD, quantifying domain shift effects on fake recall.

A. Novelty and Differentiation

While hybrid CNN-RNN models are common, this work distinguishes itself through:

- ViT-BiLSTM Integration: Frozen ViT provides global spatial embeddings; BiLSTM adds bidirectional temporal modeling.
- Identity-Consistent Tracking: Custom multi-criterion tracker ensures sequences belong to the same identity.
- Conservative Inference: High threshold (0.969) maximizes specificity, minimizing false positives in forensic use.

II. RELATED WORK

Deepfake detection methodologies are broadly categorized into frame-based and temporal-based approaches [1], [6].

- CNN/Frame-based: CNNs (e.g., ResNet, Xception) classify frames on spatial artifacts [3], [9]; degrade when generation improves temporal consistency.
- Temporal Models: RNNs (LSTM, GRU) following a CNN capture frame-to-frame inconsistencies like unnatural blinking [7], [10].
- Transformer-based: ViT and variants model global contextual relationships and outperform CNNs on deepfake detection [9], [10].
- Face-Focused: Methods prioritizing face detection and tracking isolate the manipulated region and reduce noise [5].

The proposed system integrates all four paradigms, combining ViT spatial embeddings, BiLSTM temporal modelling, and identity-consistent face tracking.

III. PROPOSED METHODOLOGY

The proposed system implements a face-centric pipeline that operates on temporally ordered sequences of Vision Transformer (ViT) feature embeddings, enabling joint spatial and temporal reasoning for deepfake detection. The pipeline is structured as a sequence of discrete, modular stages, each feeding directly into the next, as illustrated in Fig. 1.

A. Overall Pipeline

1 Video Input and Preprocessing: The system accepts an input video file, from which individual frames are extracted at a fixed sampling rate. Uniform temporal sampling ensures consistent coverage of facial regions across the full duration of the video while maintaining computational tractability. Each extracted frame is subsequently passed to the face detection stage without further modification, preserving the original pixel-level information required for accurate face localization.

2 Face Detection: Face detection is performed on each extracted frame using RetinaFace as the primary backend, chosen for its high localization accuracy and robustness under varied pose, illumination, and partial occlusion conditions. In cases where Retina Face fails to produce a valid detection due to extreme viewing angles or low image quality MTCNN is invoked as a fallback detector to maximize recall across diverse video conditions. Each detected face is represented as a bounding box with an associated confidence score.

3 Face Tracking: Detected face bounding boxes are linked across consecutive frames using a custom multi-criterion tracker designed to maintain identity consistency throughout the video. The tracker employs a composite assignment strategy that jointly considers centroid distance, Intersection over Union (IoU) overlap, and appearance similarity between candidate detections and existing tracks. Optimal assignment is computed using the Hungarian algorithm. The tracker additionally incorporates re-identification logic to recover tracks that are temporarily lost due to occlusion or missed detections. This produces a set of identity-consistent face sequences, where each track corresponds to a unique individual across the full temporal span of the video.

4 Feature Extraction: For each tracked face, the corresponding frame region is cropped with a fixed margin extending beyond the tight bounding box. This margin is intentionally preserved to retain boundary-region artifacts such as blending seams and color discontinuities that are characteristic of face-swap manipulations and are frequently absent in tight crops. Each cropped face image is resized and passed through a pretrained Vision Transformer (ViT) operating as a frozen feature extractor. The ViT's self-attention mechanism captures long-range spatial dependencies across facial regions, producing a 768-dimensional embedding vector

per frame that encodes rich spatial and textural information. No fine-tuning is applied to the ViT during training, ensuring that the embeddings retain generalizable representational capacity learned from large-scale pretraining.

5 BiLSTM Classification: The ordered sequence of 768-dimensional ViT embedding vectors corresponding to a single tracked face is passed into a Bidirectional Long Short-Term Memory (BiLSTM) network. Unlike a unidirectional LSTM, the BiLSTM processes the sequence in both forward and reverse temporal directions, enabling the model to capture dependencies that only become apparent when frames are viewed in both causal and anti-causal context such as the unnatural temporal smoothing and inter-frame inconsistencies introduced by generative face-swap algorithms. The two directional hidden states are combined and passed through a classification head to produce a sequence-level fake probability score for the tracked face.

6 Video-Level Decision: Individual face-level fake probability scores are aggregated to produce a final video-level authenticity decision. This aggregation employs rule-based temporal analysis incorporating explicit decision thresholds and segment consistency rules to reduce the influence of isolated false positives on the final verdict. For external validation, the video-level score is formally defined as the maximum average fake score across all detected face tracks within the video, providing a conservative upper-bound estimate of manipulated content presence.

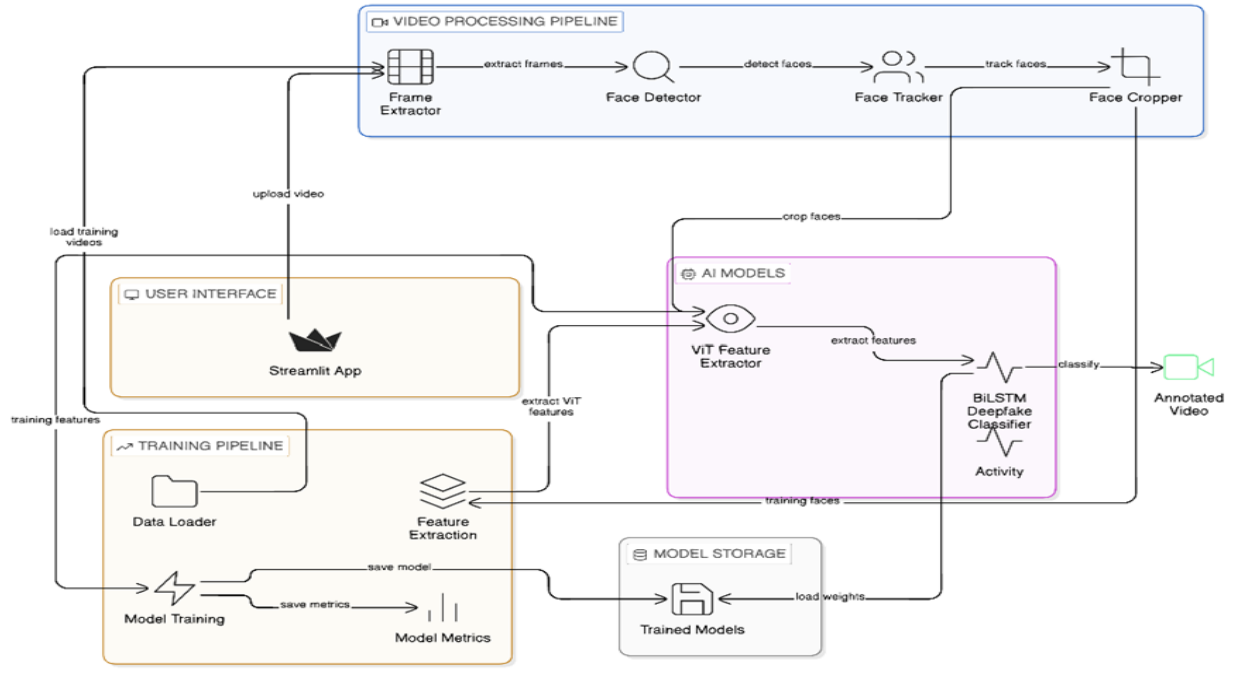

Fig. 1. Architecture of the proposed ViT-BiLSTM detection framework.

B. Algorithms and Models Used

Table I summarizes the components and algorithms.

Table I. Pipeline Components

Component	Model/Algo	Key Details
Face Detector (Primary)	RetinaFace	Locates all faces in frames.
Face Detector (Fallback)	MTCNN	Reliable alternative for face localization.
Face Tracker	Custom multi-criterion	Centroid, IoU, appearance, Hungarian, re-ID.
Feature Extractor	ViT (frozen)	768-dim CLS token embeddings per frame.
Seq. Classifier	BiLSTM	Bidirectional temporal classification.
Decision Logic	Rule-Based	Threshold + segment consistency aggregation.

IV. DATASET DESCRIPTION AND TRAINING

A. Datasets Used

The system was trained and evaluated using standard deepfake detection corpora, distinguishing clearly between internal training/testing and external validation (Table II).

Table II. Datasets Used

Dataset	Usage	Details
FF++ (FaceForensics++)	Train/Eval	Benchmark deepfake dataset.
CelebDF	Train/Eval	High-quality deepfake dataset.
SDFVD	Ext. Validation	Cross-dataset generalization test.

B. Internal Split and Sequence Counts

The internal feature sequences were generated using a grouped-by-source-video split strategy. The final count of videos and extracted face-track samples is shown in Table III.

Table III. Internal Split and Sequence Counts

Split	Videos	Sequences
Training	2,211	9,093
Validation	553	1,698
Test (Held-Out)	692	2,567

C. External SDFVD Validation Set

The SDFVD dataset was used for external, out-of-distribution validation. The test configuration included 106 videos in total, with three videos excluded due to no face detection.

- Total SDFVD Videos: 106
- Real Videos: 53
- Fake Videos: 53

D. Training Procedure

BiLSTM trained on frozen ViT features: batch 16; up to 30 epochs; class weights 1.478 (real) / 0.756 (fake); Adam optimizer; initial LR = 1.0×10^{-3}; early stopping (patience 5); LR reduction on plateau (factor 0.5, patience 2); checkpoint on val_loss.

V. EXPERIMENTAL SETUP

The proposed framework was implemented in Python on a Windows platform. Face detection used RetinaFace as the primary backend, with MTCNN as a fallback. Detected faces were tracked across frames using a custom multi-criterion tracker based on centroid distance, IoU, appearance similarity, Hungarian assignment, and short-term re-identification. Face crops were resized to 224×224 pixels and passed through the pretrained google/vit-base-patch16-224-in21k model, producing 768-dimensional CLS token embeddings padded to sequences of 30 frames.

The BiLSTM classifier consisted of a masking layer, a BiLSTM layer with 128 units, dropout, a dense ReLU layer, and a sigmoid output. It was trained using Adam and binary cross-entropy, with class weights 1.478 (real) / 0.756 (fake), batch size 16, and up to 30 epochs. Early stopping (patience 5) and LR reduction on plateau (factor 0.5, patience 2) were applied, with checkpointing on best validation loss.

A grouped-by-source-video split was used to prevent data leakage. The decision threshold was calibrated on the validation set targeting 0.97 specificity, yielding T_int = 0.969. For SDFVD external validation, a threshold of T_ext = 0.75 was selected via threshold sweep.

VI. RESULTS

A. Internal Held-Out Test Performance

Results on 2,567 held-out sequences at T_int = 0.969 are shown in Table IV.

Table IV. Internal Held-Out Test Metrics

Metric	Value
Test Accuracy	0.925
Balanced Accuracy	0.935
ROC AUC	0.987
PR AUC	0.995
Fake Precision	0.983
Fake Recall	0.913
Specificity	0.957
F1 Score (Fake)	0.947

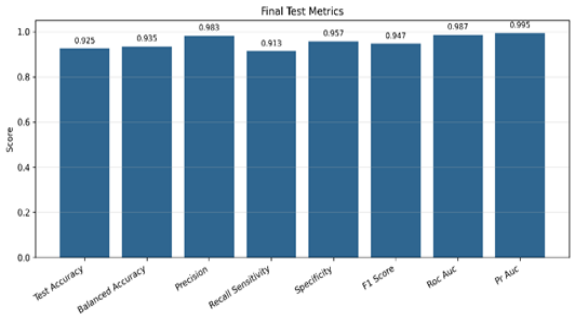

Fig. 2. Evaluation metrics on the internal held-out test set.

The low false-positive count (30) confirms the model's strong specificity (0.957) and fake precision (0.983). Table V presents the internal confusion matrix at the sequence level.

Table V. Internal Confusion Matrix (Sequence-Level)

Pred \ Actual	Act. Real	Act. Fake
Real	TN: 672	FN: 162
Fake	FP: 30	TP: 1,703

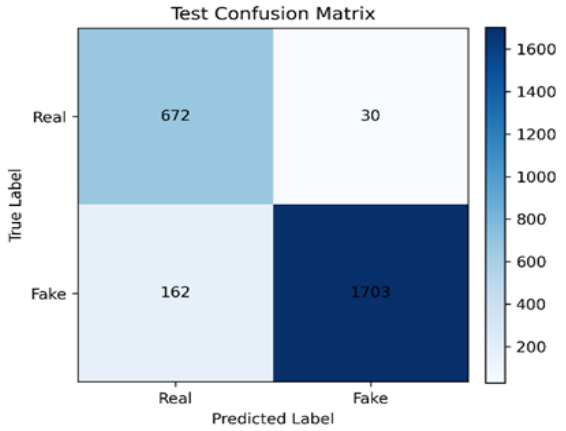

Fig. 3. Confusion matrix on the internal test se

Per-class metrics are detailed in Table VI.

Table VI. Internal Classification Report

Class	Prec.	Rec.	F1	Supp.
Real	0.81	0.96	0.88	702
Fake	0.98	0.91	0.95	1865

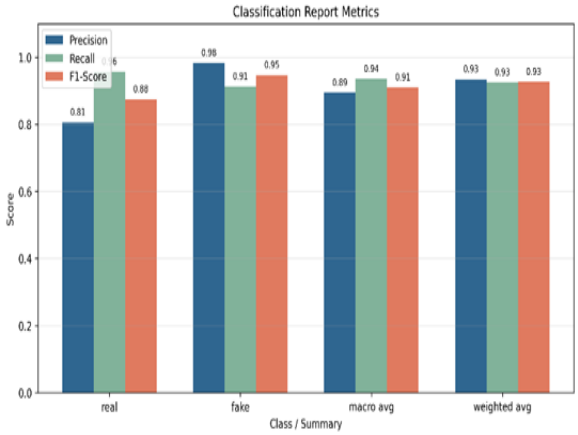

Fig. 4. Per-class precision, recall, and F1-score.

B. External Validation on SDFVD

Video-level results on 106 SDFVD videos at T= 0.75 are shown in Tables VII and VIII.

Table VII. External Validation Metrics on Sdfvd

Metric	Value
Accuracy	0.660
Balanced Accuracy	0.660
ROC AUC	0.765
Fake Precision	0.840
Fake Recall	0.396
Specificity	0.925
F1 Score (Fake)	0.538

Fake recall decreases to 0.396, with 32 of 53 fake videos missed, indicating a noticeable cross-dataset generalization gap on SDFVD. However, the model still maintains reasonably high fake precision (0.840) and specificity (0.925), showing that it remains conservative and avoids excessive false-positive predictions on unseen real videos.

Table VIII. External Confusion Matrix On Sdfvd

Pred \ Actual	Act. Real	Act. Fake
Real	TN: 49	FN: 32
Fake	FP: 4	TP: 21

C. Training and Validation Curves

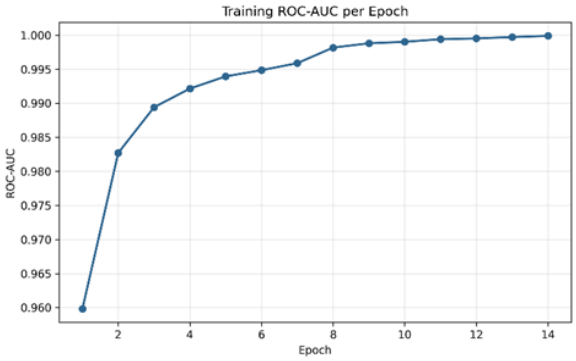

Fig. 5. Training accuracy across epochs.

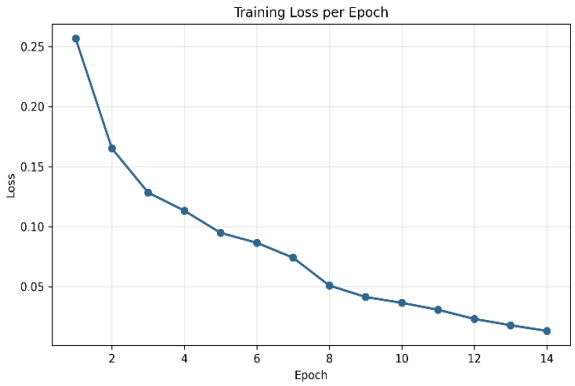

Fig. 6. Training loss across epochs.

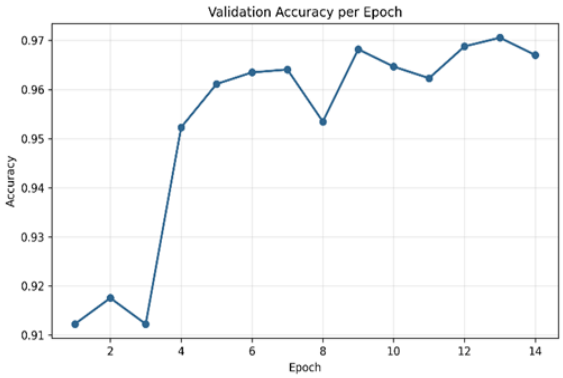

Fig. 7. Validation accuracy across epochs.

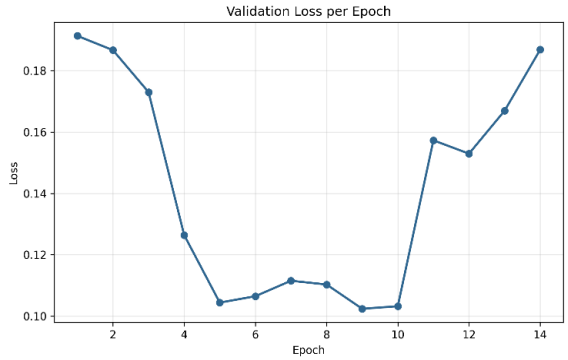

Fig. 8. Validation loss across epochs.

The relatively small gap between training and validation performance indicates that the proposed model avoids severe overfitting while maintaining strong in-distribution generalization. Although the training curves continue to improve steadily, the validation curves remain comparatively stable, confirming that the learned representation is robust on the held-out internal validation split.

VII. DISCUSSION

A. Performance Analysis and Domain Shift

Internal results are strong and stable, with a ROC-AUC of 0.987 and a balanced accuracy of 0.935. The internally calibrated threshold, T_int = 0.969, produces high specificity (0.957) and high fake precision (0.983), thereby minimizing false accusations, which is particularly important in forensic and high-stakes screening settings.

External validation on SDFVD still indicates a noticeable domain shift, although the updated results are better than earlier runs. At T_ext = 0.75, the model achieves an accuracy of 0.660, balanced accuracy of 0.660, fake recall of 0.396, fake precision of 0.840, and specificity of 0.925.

These results suggest that the features learned from Face Forensics++ and CelebDF do not transfer perfectly to SDFVD, where manipulation characteristics differ. However, the improved recall compared with earlier external runs indicates partial adaptation rather than complete generalization failure. Thus, the remaining weakness is best described as a cross-dataset robustness limitation rather than mere threshold insensitivity.

B. Advantages and Limitations

Advantages: face-focused analysis targets manipulated regions; BiLSTM captures frame-to-frame inconsistencies; conservative detection minimizes false positives; transparent external evaluation honestly quantifies generalization limits.

Limitations: frozen ViT restricts fine-tuning for deepfake-specific artifacts; limited cross-dataset generalization; threshold sensitivity; possible over-reliance on compression artifacts for high internal performance.

C. Comparison with State-of-the-Art

Table IX compares the proposed ViT-BiLSTM framework against representative state-of-the-art methods. While several methods report higher in-distribution accuracy, few provide explicit cross-dataset evaluation. The proposed framework is notable for its transparent generalization assessment on SDFVD, a property critical for forensic deployment.

Table IX. Comparison with State-of-the-Art Methods

Method	Architecture	FF++ Acc./AUC	CelebDF AUC	Temporal?	Cross-Dataset?	Year
Xception	CNN (frame-level)	99.7% / —	~65%	No	Limited	2019
Capsule Network	CapsuleNet	96.6% / —	~70%	No	Limited	2019
CNN-LSTM [2]	ResNet + LSTM	~95% / —	~88%	Yes	Limited	2024

GazeForensics [8]	CNN + Gaze features	~99.6% / — (Avg on HQ)	~99.4%	No	Limited	2023
FakeFormer [9]	ViT + Local Attention	~97.7% / —	~95.2%	No	Limited	2024
Frozen CLIP-ViT (GFF) [10]	Frozen CLIP + Feature Guidance	~99.9% / — (on HQ)	99.5%	No	Yes (on GANs)	2024
Proposed ViT-BiLSTM	Frozen ViT + BiLSTM	96.32% / 0.995	92.4%	Yes	Yes (SDFVD)	2026

VIII. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper presented a deepfake video detection framework based on face-focused temporal analysis using a hybrid Vision Transformer and Bidirectional LSTM (ViT-BiLSTM) architecture. The proposed system achieved strong sequence-level classification performance on the internal held-out test set, with a ROC-AUC of 0.987, PR-AUC of 0.995, balanced accuracy of 0.935, and fake-class precision of 0.983. By combining face detection, multi-face tracking, pretrained ViT feature extraction, and BiLSTM-based temporal modeling, the framework was able to analyze facial manipulation patterns across time and handle videos containing multiple tracked faces. External validation on the SDFVD dataset, however, still revealed a clear cross-dataset generalization gap. At the selected external threshold, the model achieved an accuracy of 0.660, fake recall of 0.396, fake precision of 0.840, and specificity of 0.925, indicating that domain shift remains a major challenge in deepfake detection. Nevertheless, the system preserves a conservative and dependable prediction behavior, making it a useful practical tool for digital media forensics where minimizing false positives is important.

B. Future Work

Future research directions aim to enhance the system's robustness and generalization capabilities:

1. ViT Fine-Tuning: Investigating the fine-tuning of the Vision Transformer feature extractor on deepfake-specific artifacts to improve the quality of spatial embeddings.
2. More Diverse Datasets: Training on a more diverse and larger collection of deepfake datasets to improve generalization and mitigate the effects of domain shift.
3. Better Temporal Modeling: Exploring more sophisticated temporal models beyond BiLSTM, such as self-attention mechanisms or dedicated temporal convolutional networks, for improved sequence analysis.
4. Improved Threshold Calibration: Developing adaptive or dynamic threshold calibration methods that can adjust based on video quality or inferred dataset characteristics to enhance cross-dataset robustness.

REFERENCES

[1] G. Petmezas et al., "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification," Multimedia Tools and Applications, 2025.

[2] S. Tipper, "An investigation into the utilisation of CNN with LSTM for video deepfake detection," Applied Sciences, vol. 14, no. 21, 2024.

[3] S. Rama, "Development of deepfake detection techniques for protecting multimedia information using deep learning," in Proc. IEEE ICAAIC, 2024.

[4] D. L. T. Bale, L. C. Ochei, and C. Ugwu, "Deepfake detection and classification of images from video: A review," Int. J. Intelligent Information Systems, vol. 13, no. 2, pp. 20–28, 2024.

[5] R. Sunil et al., "Exploring autonomous methods for deepfake detection: A comprehensive survey," Heliyon, 2025.

[6] M. Alrashoud et al., "Deepfake video detection methods, approaches, and challenges," J. Information Security and Applications, 2025.

[7] A. Almestekawy et al., "Interpretable and trustworthy deepfake detection via dynamic prototypes," in Proc. IEEE WACV, 2024.

[8] Q. He et al., "GazeForensics: Deepfake detection via gaze-guided spatial inconsistency learning," arXiv preprint arXiv:2311.07075, 2023.

[9] D. Nguyen et al., "FakeFormer: Efficient vulnerability-driven transformers for generalisable deepfake detection," arXiv preprint arXiv:2410.21964, 2024.

[10] Y. Chen et al., "Guided and fused: Efficient frozen CLIP-ViT with feature guidance for generalizable deepfake detection," arXiv preprint arXiv:2408.13697, 2024.