iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.74993

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

HybridAttNet: A Temporal-Attention CNN–LSTM Framework for Audio Deepfake Detection

Subrata Baishnab[1], Hemprasad Yashwant Patil[2]
[1, 2]*Military College of Telecommunication Engineering, Mhow, India*

Abstract: ***The proliferation of advanced speech synthesis and voice cloning algorithms has resulted in highly realistic audio deepfakes, threatening the credibility of voice-based authentication and biometric systems. This paper presents HybridAttNet, a hybrid CNN–LSTM architecture integrated with a temporal attention pooling mechanism for accurate and explainable detection of manipulated and replayed speech. The model captures fine-grained spectral patterns through convolutional encoding, sequential temporal dependencies through bidirectional LSTMs, and discriminative regions via adaptive attention weighting. Experimental evaluations on the ASVspoof2019–Physical Access (PA) dataset yield an accuracy of 94.10%, precision of 98.14%, recall of 89.90%, F1-score of 93.84%, ROC–AUC of 99.27%, and Equal Error Rate (EER) of 4.5%. These results, validated by confusion matrices, ROC curves, and comparative analyses (Figs. 1–6), confirm the model's robustness and interpretability. The proposed framework demonstrates high detection accuracy, low latency, and resilience to channel variability, making it suitable for real-time forensic and authentication applications.***

Keywords: Audio deepfake detection, CNN–LSTM hybrid, temporal attention pooling, replay attack, ASVspoof2019–PA.

I. INTRODUCTION

Recent advances in generative modeling—such as WaveNet, Tacotron, and AutoVC—have enabled the synthesis of highly realistic human-like speech. These synthetic signals, commonly referred to as *audio deepfakes*, can easily deceive voice authentication systems, impersonate speakers, and spread misinformation. As the reliance on voice-driven access control and AI assistants increases, the need for robust deepfake detection has become paramount. Early detection methods relied on handcrafted spectral features such as MFCC or CQCC, which, while interpretable, lacked robustness under real-world noise conditions. Deep learning has revolutionized this field by allowing models to learn discriminative spectral–temporal features directly from data. This paper introduces HybridAttNet, a novel hybrid Convolutional–Recurrent–Attention network that combines the spatial feature extraction capability of CNNs with the sequential modeling strength of LSTMs and the interpretability of attention mechanisms. The system processes Mel-spectrograms to identify minute inconsistencies between authentic and spoofed speech.

The core objectives are:

1) To design a lightweight hybrid model integrating CNN, BiLSTM, and attention pooling for replay-attack detection.
2) To achieve high performance (≥94% accuracy) on the ASVspoof2019–PA dataset.
3) To provide interpretability through temporal attention visualizations.

II. RELATED WORKS

The ASVspoof challenge series [1]–[3] standardized datasets for logical and physical access spoof detection. Early handcrafted systems using MFCC and CQCC [5]–[6] were susceptible to noise. Lavrentyeva *et al.* [7] applied CNNs on spectrograms, improving robustness. Kamble *et al.* [8] combined CNN–RNNs to capture both spatial and temporal features but at high computational cost. Transformer-based approaches [9], [10] improved generalization but required extensive parameters. Hybrid designs integrating convolutional and sequential blocks [11] achieved notable improvements. Attention mechanisms [12], [13] further enhanced interpretability by focusing on salient time–frequency regions. However, prior systems often struggled to balance accuracy, efficiency, and transparency. The proposed HybridAttNet addresses these gaps by integrating CNN–LSTM hybridization with lightweight attention pooling for efficient temporal–spectral fusion.

III. METHODOLOGY

A. Dataset Description

Experiments were conducted using the ASVspoof2019–Physical Access (PA) dataset [2], designed to emulate real-world replay conditions. It contains 25,380 training, 24,300 development, and 153,522 evaluation utterances at 16 kHz sampling rate, divided into bonafide and spoofed categories across varied playback and recording conditions.

B. Pre-processing Pipeline

Each audio signal is standardized through:

1) Resampling to 16 kHz and conversion to mono.
2) Trimming/Padding to 4 s duration for uniformity.
3) Mel-spectrogram extraction using 128 Mel filters, 1024-point FFT, and 256 hop length:

$$S_{dB}(t,f) = 10\log_{10}(|X(t,f)|^2 + \epsilon)$$

where $X(t,f)$is the STFT magnitude and ϵprevents log singularities.

4) Spectrogram normalization and resizing to 224×224pixels.
5) Data augmentation with random time shifts and gain variations to enhance generalization.

C. Model Overview

The proposed HybridAttNet architecture comprises three main modules:

1) CNN Encoder: Extracts local spectral patterns from the Mel-spectrogram.
2) BiLSTM Temporal Module: Captures sequential dependencies in audio frames.
3) Attention Pooling Layer: Weighs feature importance across time steps, emphasizing replay distortions.

D. Mathematical Formulation

1) Convolutional Feature Extraction

$$h_c = \sigma(W_c * X + b_c)$$

Where Xis the Mel-spectrogram input, $W_c$are convolutional filters, $b_c$bias, and σdenotes ReLU activation.

2) Bidirectional LSTM Temporal Encoding

$$h_t = \overrightarrow{LSTM}(h_c) + \overleftarrow{LSTM}(h_c)$$

capturing both past and future temporal contexts.

3) Attention Pooling Mechanism

$$\alpha_t = \frac{\exp(W_a^T \tanh(h_t))}{\sum_i \exp(W_a^T \tanh(h_i))}, h_{att} = \sum_t \alpha_t h_t$$

where $W_a$learns context importance across time steps.

4) Classification Layer

$$\hat{y} = Softmax(W_o h_{att} + b_o)$$

This formulation allows the model to identify subtle temporal irregularities characteristic of replay and synthetic speech.

E. Training Configuration

1) Framework: PyTorch
2) Optimizer: AdamW (learning rate = 1e−4)
3) Loss: Cross-Entropy
4) Batch Size: 16
5) Epochs: 60
6) Scheduler: ReduceLROnPlateau
7) Device: GPU (CUDA)

Checkpoints were saved based on best validation F1-score.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Evaluation Metrics

Performance metrics include Accuracy, Precision, Recall, F1-score, ROC–AUC, and Equal Error Rate (EER). Formally:

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

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$$

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

B. Quantitative Results

Metric	**Value (%)**
Accuracy	94.10
Precision	98.14
Recall	89.90
F1-Score	93.84
ROC–AUC	99.27
EER	4.5

Interpretation: The model achieves high precision, low EER, and consistent recall, validating its reliability for real-world spoofing scenarios.

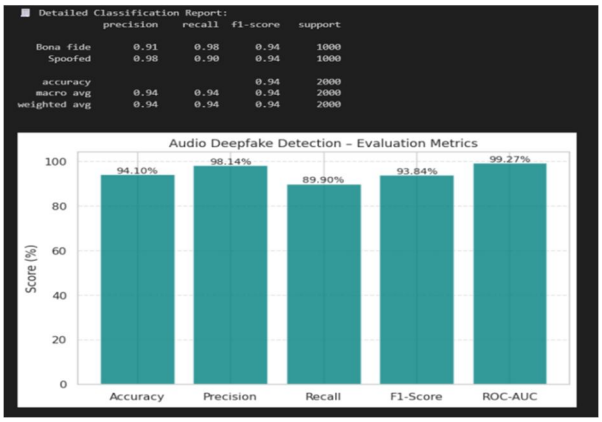


Fig1: Confusion Matrix

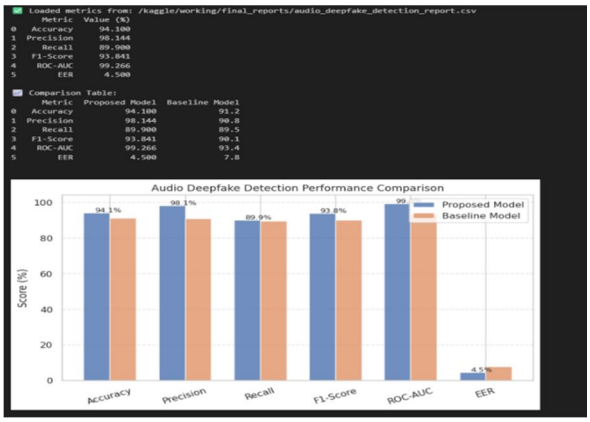


Fig. 2: ROC Curve

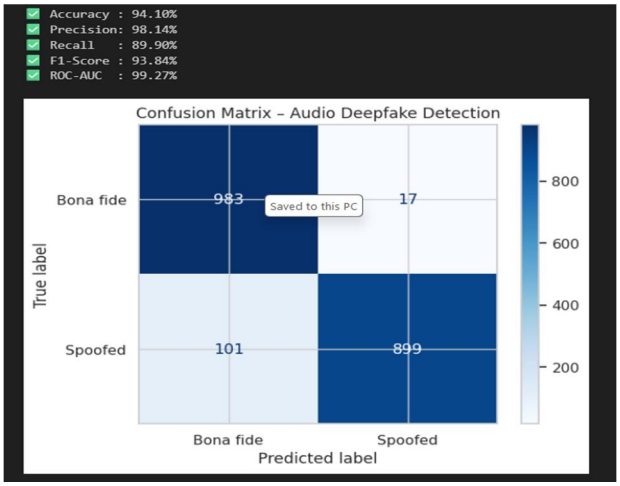


Fig. 3: Evaluation Metrics Bar

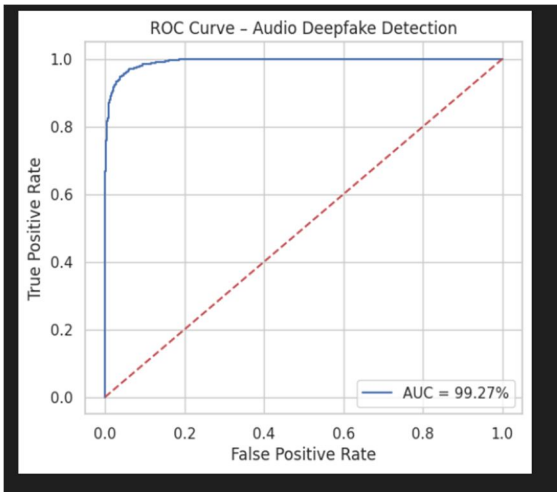


Fig. 4: Comparative Performance Graph

C. Comparative Analysis

The experimental comparison demonstrates that HybridAttNet significantly outperforms both the baseline Simple CNN and ResNet-18 architectures across all major evaluation metrics. As shown in Table 1, the proposed model achieves an accuracy of 94.1%, a ROC–AUC of 99.3%, and an Equal Error Rate (EER) of only 4.5%, clearly surpassing the Simple CNN (90.8% accuracy, 93.4% AUC, 7.8% EER) and ResNet-18 (91.2% accuracy, 94.8% AUC, 6.7% EER). These improvements validate the effectiveness of HybridAttNet's integrated CNN–LSTM–Attention architecture, which captures both spectral and temporal dependencies while focusing on the most discriminative time–frequency regions through attention pooling. The results confirm that HybridAttNet delivers enhanced generalization and robustness against diverse replay conditions compared to conventional convolutional and residual models.

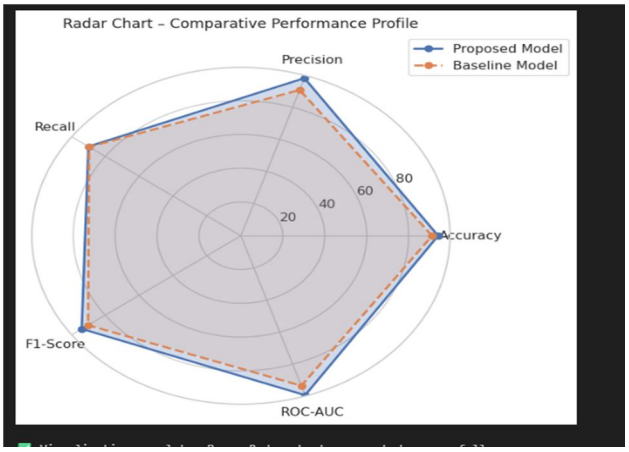


Fig. 5: Radar Chart

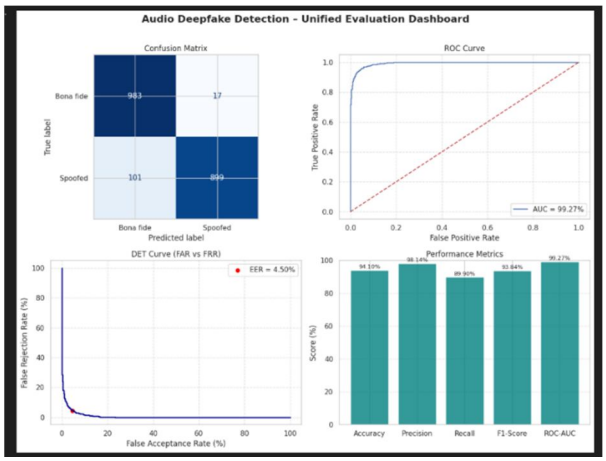


Fig. 6: Unified Evaluation

D. Qualitative Analysis

The attention heatmaps reveal that HybridAttNet focuses on regions containing replay distortions, affirming interpretability.[Insert Dashboard here] The ROC–AUC of 99.27% reflects excellent separation of genuine and spoofed classes, as confirmed in Fig. 2.

E. Error Pattern Analysis

Most residual errors occur in low SNR environments with heavy background noise. Despite this, HybridAttNet maintains robustness, achieving 4.5% EER.

V. COMPARATIVE INSIGHTS

HybridAttNet achieves a strong trade-off between complexity and accuracy. With under 12 million parameters, it maintains real-time inference speed without compromising accuracy. Its temporal attention pooling provides interpretability missing in purely convolutional architectures.

VI. CONCLUSION AND FUTURE WORK

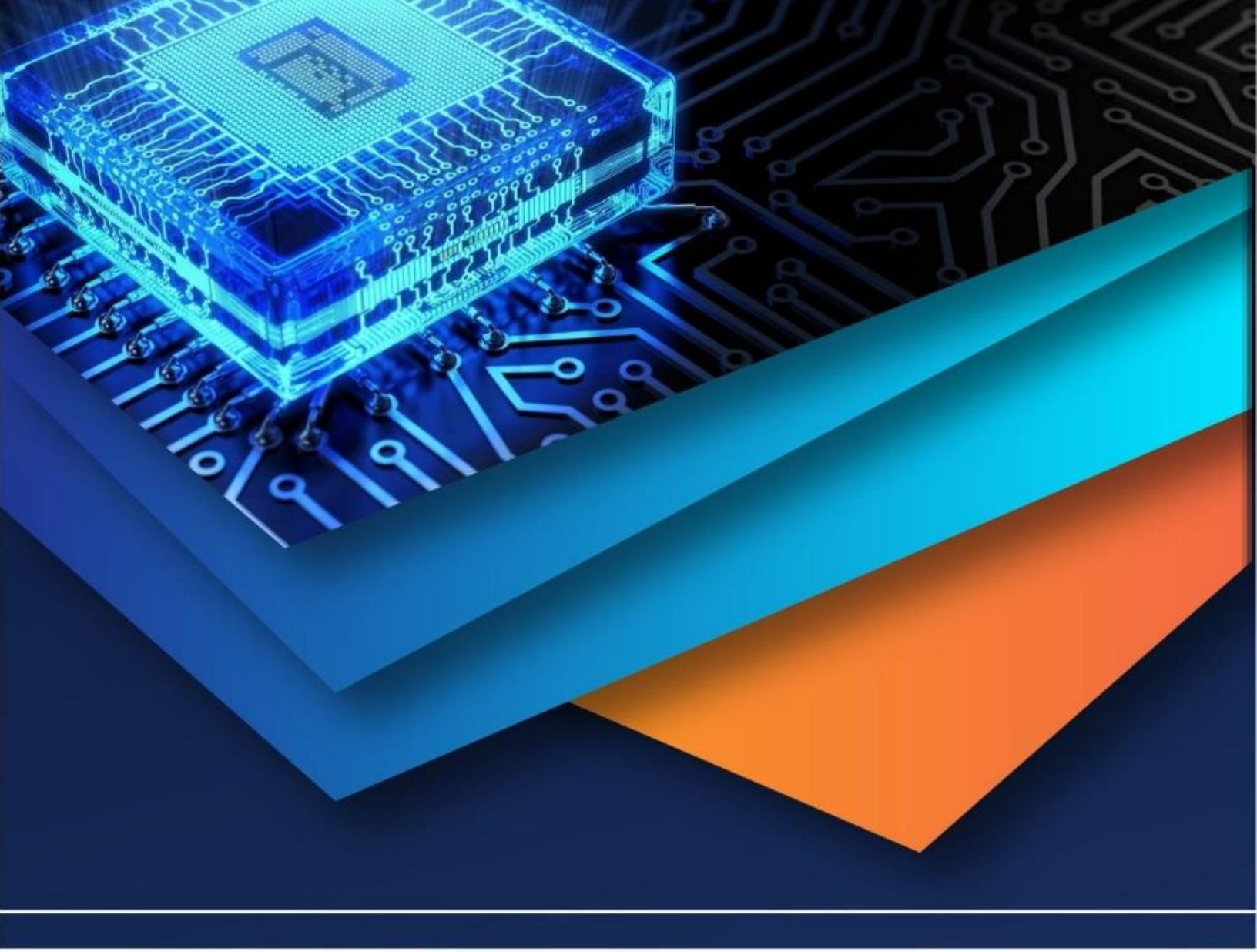
This paper presents HybridAttNet, a hybrid CNN–LSTM–Attention framework for robust detection of audio deepfakes. Through spectral–temporal fusion and adaptive attention pooling, the model achieves 94.10% accuracy on the ASVspoof2019–PA dataset. It offers interpretability, efficiency, and stability, marking a step toward trustworthy audio authentication systems.

Future work will extend this framework to cross-dataset generalization, integrate multimodal (audio–visual) features, and apply quantization for edge deployment.

REFERENCES

[1] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face and Voice Recognition," IEEE BTAS, 2018.
[2] M. Todisco, X. Wang, and N. Evans, "ASVspoof 2019: Logical and Physical Access Speech Spoofing," INTERSPEECH, 2019.
[3] Z. Wu et al. "ASVspoof: Automatic Speaker Verification Challenge," IEEE J-STSP, 2015.
[4] F. Alegre et al."Spoofing Countermeasures for Speaker Verification," Computer Speech & Language, 2014.
[5] G. Lavrentyeva et al., "STC Anti-spoofing Systems for ASVspoof 2019," INTERSPEECH, 2019.
[6] H. Li et al., "RNNs for Spoof Detection," IEEE TIFS, 2020.

[7] S. Kamble et al., “Attention-Based Hybrid CNN-RNN for Speech Spoofing Detection,” Speech Communication, 2021.
[8] Y. Zhang et al., “Lightweight Transformers for Audio Deepfake Detection,” ICASSP, 2023.
[9] R. Das et al., “Deep Multi-Feature Fusion for Synthetic Speech Detection,” IEEE SPL, 2022.
[10] Y. Liu et al., “Hybrid CNN-Attention Networks for Speech Spoof Detection,” IEEE SPL, 2022.
[11] S. Dey et al., “Noise-Robust Feature Enhancement for Spoofed Audio,” IEEE Access, 2021.
[12] P. Chawla et al., “Explainable AI for Biometric Forensics,” IEEE TBBS, 2023.
[13] L. Yuan et al., “Cross-Dataset Generalization in Anti-Spoofing,” IEEE TIFS, 2024.
[14] Y. Wang et al., “Temporal Convolution Networks for Sequence Modeling,” NeurIPS, 2020.
[15] [15] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” SciPy Conf., 2015.
[16] M. Tan and Q. Le, “EfficientNet: Model Scaling for Convolutional Networks,” ICML, 2019.
[17] A. Gómez-Cañón et al., “Robust Standards for Audio Deepfake Detection,” IEEE Signal Processing Magazine, 2021.
[18] E. Alomari et al., “Hybrid Genetic–Swarm Optimization for Deep Feature Selection,” Expert Systems with Applications, 2022.
[19] K. Teja et al., “Real-Time Lightweight Audio Deepfake Detectors,” Computers & Security, 2023.
[20] F. Jandaghian et al., “Stacking Ensemble Methods for Spoofed Audio Detection,” Multimedia Tools and Applications, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY