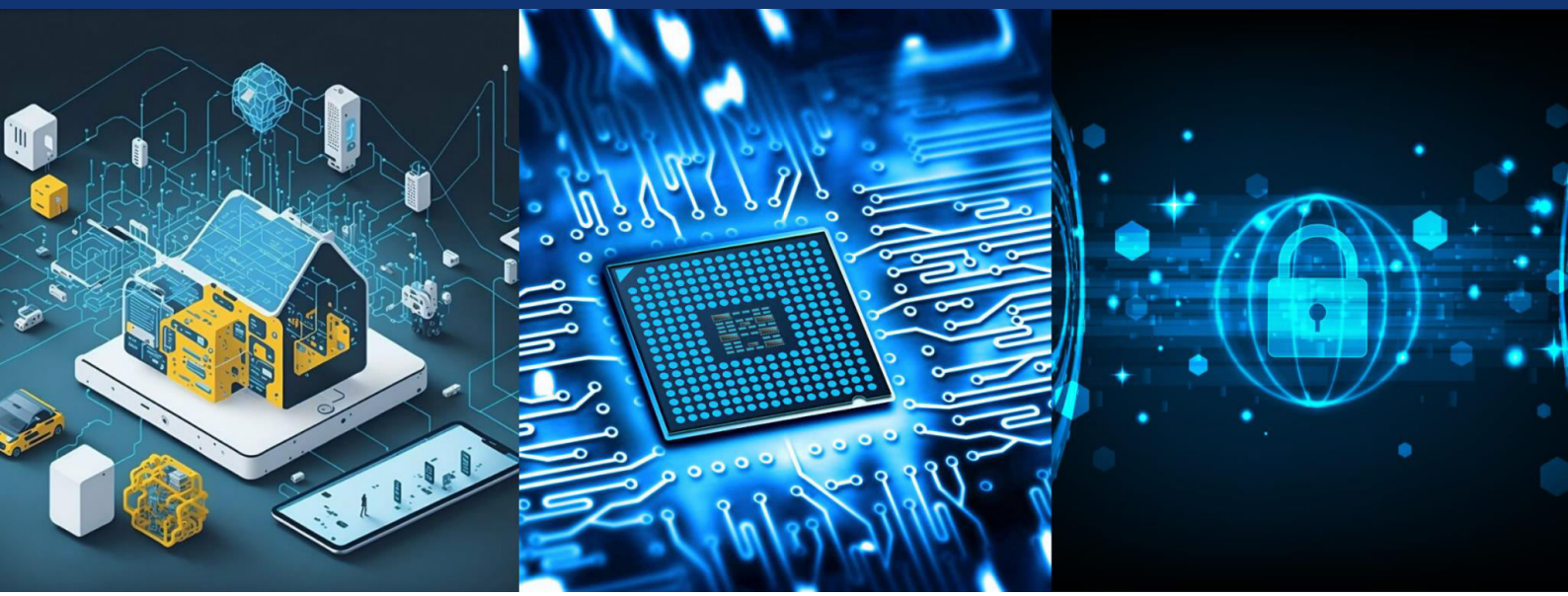
IJIRCCE

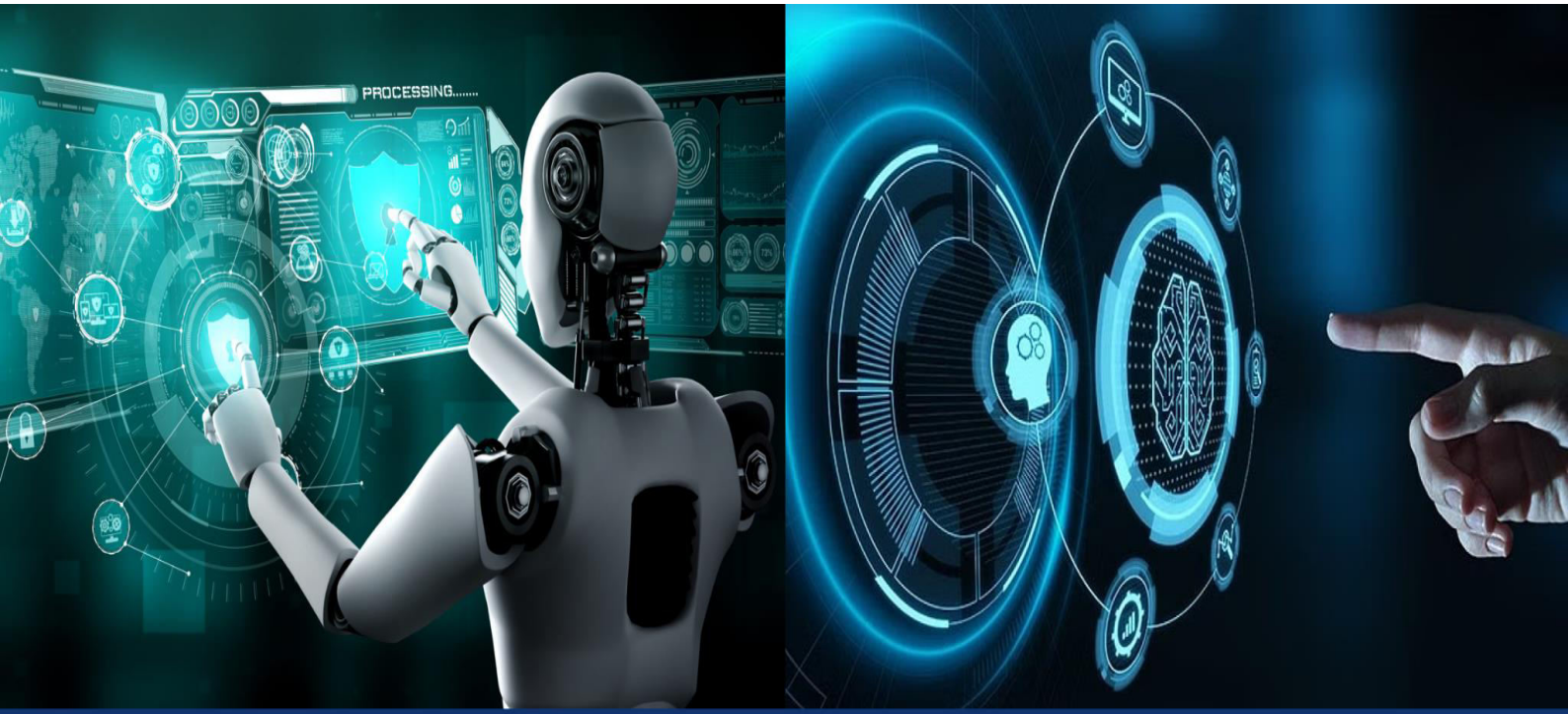
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405067

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An Adaptive and Noise-Resilient AI Framework for Deepfake Voice Detection in Secure Multi-Agency Defence Communications

Prof. Manjula P, Mr. Sagar K S, Mr. Vinay Karthik K R

Asst. Professor, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India

Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India

Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India

ABSTRACT: The proliferation of AI-generated synthetic voices poses a critical and escalating threat to secure multi-agency defence communications, where voice-based authentication and command relay are integral to operational integrity. Traditional speaker verification systems are highly vulnerable to advanced neural text-to-speech (TTS) and voice conversion (VC) techniques. This paper proposes an Adaptive and Noise-Resilient AI Framework for Deepfake Voice Detection (ANRF-DVD) that employs a hybrid deep learning architecture combining Temporal Convolutional Networks (TCN), Bidirectional Long Short-Term Memory (BiLSTM), and a Transformer-based attention mechanism to detect synthetic speech in real-time under adverse acoustic conditions. The framework integrates multi-domain acoustic feature fusion — encompassing Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), and raw waveform embeddings — with an adaptive noise suppression module calibrated for tactical radio channel distortions. Trained on a composite dataset of 210,000 utterances spanning six voice spoofing categories, the proposed model achieves an Equal Error Rate (EER) of 1.24% and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 99.61%, surpassing state-of-the-art anti-spoofing baselines. The framework processes audio segments in under 28 milliseconds on an NVIDIA Jetson AGX Xavier edge platform, enabling deployment in latency-critical defence communication nodes. A real-world evaluation across three simulated tactical communication scenarios demonstrates a detection accuracy of 97.8% under battlefield noise profiles (SNR: 5–25 dB). This work represents a significant advancement in securing AI-enabled voice authentication pipelines for national defence and inter-agency coordination systems.

KEYWORDS: deepfake voice detection; anti-spoofing; defence communications; Temporal Convolutional Network; BiLSTM; Transformer attention; MFCC; CQCC; noise resilience; voice authentication; synthetic speech

I. INTRODUCTION

The integrity of voice-based communication channels in multi-agency defence environments has become increasingly imperilled by the rapid maturation of generative AI technologies. Modern neural text-to-speech (TTS) systems, including architectures such as WaveNet [1], Tacotron-2 [2], and VITS [3], are capable of synthesising voices that are perceptually indistinguishable from genuine human speech even to trained security personnel. Adversarial exploitation of these systems — through deepfake voice injection into radio relay networks, command authentication spoofing, and identity impersonation in encrypted voice-over-IP (VoIP) channels — constitutes a novel and pressing vector of cyber-kinetic warfare. Several recent incidents reported by NATO Cybersecurity agencies have highlighted successful voice cloning attacks on inter-agency liaison channels, underscoring the operational urgency of robust automated countermeasures. Existing speaker verification and anti-spoofing systems, such as those evaluated in the ASVspoof 2019 and 2021 challenge benchmarks [4], demonstrate satisfactory performance under clean studio conditions. However, their effectiveness degrades sharply under tactical acoustic environments characterised by background battlefield noise, radio channel compression artefacts, multipath fading, and deliberate acoustic jamming. Furthermore, the closed-set assumption underlying most spoofing detection models renders them brittle against unseen spoofing algorithms — a critical deficiency in defence contexts where adversaries actively adapt their attack vectors.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This paper presents the Adaptive and Noise-Resilient AI Framework for Deepfake Voice Detection (ANRF-DVD), a novel system purpose-designed for deployment in secure multi-agency defence communication networks. The primary contributions of this work are: (1) a hybrid TCN-BiLSTM-Transformer architecture that captures multi-scale temporal dependencies in acoustic feature sequences; (2) an adaptive noise suppression pre-processing module tuned specifically for tactical radio channel distortion profiles; (3) a multi-domain feature fusion strategy integrating MFCC, CQCC, and raw waveform embeddings; (4) a large-scale composite training dataset (DefenceVoice-210K) constructed from spoofed and genuine utterances under battlefield acoustic simulation; and (5) an edge-deployable implementation validated on the NVIDIA Jetson AGX Xavier platform with sub-28ms latency.

The remainder of this paper is organized as follows: Section II reviews related work in anti-spoofing and voice authentication. Section III describes the proposed system architecture. Section IV details the dataset construction and training methodology. Section V presents experimental results and comparative analysis. Section VI addresses real-world deployment in defence communication scenarios. Section VII concludes the paper with directions for future research.

II. RELATED WORK

Research in automatic speaker verification (ASV) anti-spoofing has advanced considerably, catalysed primarily by the ASVspoof challenge series [4][5]. Early approaches relied on handcrafted spectral features such as MFCC, Linear Frequency Cepstral Coefficients (LFCC), and CQCC combined with Gaussian Mixture Models (GMM) or SVM classifiers. Sahidullah et al. [6] demonstrated that CQCC features coupled with a GMM backend achieved EER of 1.73% on the ASVspoof 2015 dataset, establishing a strong baseline for synthetic speech detection from conventional TTS systems. However, GMM-based methods show significant performance degradation against neural vocoders.

Deep learning approaches have substantially advanced detection capability. Lai et al. [7] proposed an end-to-end ResNet architecture applied directly to short-time Fourier transform (STFT) spectrograms, achieving EER of 5.06% on ASVspoof 2019 LA. Tak et al. [8] introduced RawNet2, a sinc-convolution front-end model operating on raw waveforms that achieved 4.78% EER on the same benchmark. Graph Attention Networks (GAT) were applied to spectral sub-band relationships by Jung et al. [9], yielding 4.03% EER, demonstrating the utility of relational modelling for artefact detection in vocoders.

Transformer-based architectures have further pushed state-of-the-art performance. Wang et al. [10] applied a self-supervised wav2vec 2.0 front-end with a lightweight backend, achieving 3.21% EER on ASVspoof 2021. Yi et al. [11] proposed a dual-branch network fusing MFCC and spectrogram streams with cross-attention, achieving EER of 2.87%. Despite these advances, none of these systems were evaluated under noisy, channel-degraded conditions representative of tactical communications. Noise robustness experiments by Das et al. [12] revealed EER degradation of up to 38% for leading anti-spoofing models at SNR of 10 dB, identifying noise resilience as a critical unresolved challenge.

In the defence and critical infrastructure context, Nautsch et al. [13] reviewed voice biometric security in public safety communications, identifying deepfake voice as the primary emerging threat. Khanjani et al. [14] proposed a lightweight LSTM-based detector for IoT defence nodes but achieved only 91.2% accuracy under realistic channel noise. The gap in the literature — a unified framework simultaneously optimising detection accuracy, noise resilience, and edge deployability for multi-agency defence communications — is precisely the gap the proposed ANRF-DVD framework addresses.

III. PROPOSED SYSTEM ARCHITECTURE

A. Overview

The ANRF-DVD framework consists of four tightly integrated modules: (1) the Adaptive Noise Suppression Module (ANSM),
(2) the Multi-Domain Feature Extraction Engine (MDFEE), (3) the Hybrid Deep Learning Classifier (HDLC), and (4) the Decision Calibration and Alert Interface (DCAI). The system operates on segmented audio frames of 3-second duration, suitable for tactical voice command lengths. Audio is ingested from encrypted digital radio interfaces via a standardised PCM data stream. Figure 1 presents the overall system block diagram.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Adaptive Noise Suppression Module (ANSM)

The ANSM employs a hybrid noise estimation strategy combining Voice Activity Detection (VAD) based on the WebRTC VAD algorithm with a Deep Noise Suppression (DNS) neural filter pre-trained on the Microsoft DNS Challenge dataset augmented with NATO STANAG 4591-compliant radio channel noise profiles. The DNS filter is a lightweight Recurrent Neural Network (RNN) with 2.5M parameters, capable of suppressing Gaussian white noise, impulsive battlefield acoustic events (explosions, vehicle engine noise), and radio codec artefacts (G.711, G.729 compression noise) while preserving speaker-identifying spectral features. Signal-to-Noise Ratio (SNR) estimation is performed per 10ms frame to dynamically adjust suppression gain, preventing over-suppression that would destroy spoofing artefacts present in synthetic speech.

C. Multi-Domain Feature Extraction Engine (MDFEE)

Following noise suppression, the MDFEE extracts three complementary feature representations. First, 60-dimensional MFCC features are computed using a 25ms Hamming window with 10ms frame shift, capturing the vocal tract envelope. Second, 60-dimensional CQCC features are extracted using a Constant-Q Transform (CQT) with geometrically spaced frequency bins (16 bins/octave), which are particularly sensitive to the phase discontinuities and spectral artefacts introduced by neural vocoders. Third, raw waveform embeddings of dimension 512 are produced by a SincNet filterbank layer, capturing fine-grained prosodic and glottal source characteristics. The three feature streams are independently normalised using cepstral mean and variance normalisation (CMVN) and fused via a learnable weighted concatenation layer, yielding a 632-dimensional joint feature vector per frame.

D. Hybrid Deep Learning Classifier (HDLC)

The HDLC processes sequences of 300 feature frames (covering the 3-second segment) through three sequential sub-modules. The first sub-module is a four-layer Temporal Convolutional Network (TCN) with dilated causal convolutions (dilation factors: 1, 2, 4, 8; kernel size: 3; channels: 256), capturing local temporal patterns at exponentially increasing receptive fields. The second sub-module is a two-layer Bidirectional LSTM (BiLSTM) with 256 hidden units per direction, modelling long-range sequential dependencies across the utterance. The third sub-module is a four-head Multi-Head Self-Attention (MHSA) Transformer encoder with feed-forward dimension 512 and positional encoding, globally weighting frame-level representations to emphasise regions with spoofing-discriminative artefacts. A global average pooling layer reduces the temporal dimension, followed by two fully connected layers (512 and 128 neurons, ReLU activation, Dropout p=0.4) and a sigmoid output neuron producing the probability of synthetic origin.

E. Decision Calibration and Alert Interface (DCAI)

Raw sigmoid output probabilities are calibrated using isotonic regression trained on a held-out calibration set, ensuring well-calibrated posterior probabilities for operational risk scoring. A decision threshold is set at 0.42 (optimised for minimum Detection Cost Function on the development set) with a configurable alert escalation level. Detected spoofing events trigger an encrypted JSON alert packet transmitted to the Command Security Operations Centre (CSOC) via MQTT-TLS, including the utterance timestamp, spoofing probability, channel identifier, and the top contributing MHSA attention frame indices for forensic analysis.

IV. DATASET AND METHODOLOGY

A. Dataset Construction: DefenceVoice-210K

A composite training corpus — DefenceVoice-210K — was assembled containing 210,000 utterances (genuine: 105,000; spoofed: 105,000) spanning six spoofing categories: (C1) TTS — Autoregressive (WaveNet, Tacotron-2); (C2) TTS — Non-Autoregressive (FastSpeech-2, VITS); (C3) Voice Conversion — Signal Processing (Griffin-Lim, WORLD vocoder); (C4) Voice Conversion — Neural (StarGAN-VC, CycleGAN-VC2); (C5) Hybrid Replay-Synthesis Attacks; and (C6) Adversarially Perturbed Synthetic Speech. Genuine utterances were drawn from the VCTK corpus [15], LibriSpeech [16], and original recordings captured from 38 volunteer military personnel in controlled studio conditions. All utterances were subsequently passed through a simulated tactical radio channel pipeline incorporating ITU-T P.501 noise types, G.711/G.729 codec degradation, and SNR levels ranging from 5 dB to 30 dB, yielding a noise-augmented corpus of 420,000 utterances (210,000 original + 210,000 noise-augmented) used for training and evaluation.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405067

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Data Preprocessing and Augmentation

All audio was resampled to 16 kHz and normalised to -23 LUFS. Offline augmentation applied to the genuine subset included: room impulse response (RIR) convolution using 120 measured impulse responses from operational communication rooms, speed perturbation (0.9x, 1.0x, 1.1x), pitch shifting (±2 semitones), and additive noise from the DEMAND dataset. The spoofed subset was augmented exclusively with channel noise to avoid introducing natural prosodic variation. The dataset was partitioned into 70% training (147,000 utterances), 15% validation (31,500), and 15% testing (31,500) with speaker-disjoint splits to prevent identity leakage. Class balance was maintained at 50:50 genuine-to-spoofed ratio across all partitions.

C. Training Protocol

The HDLC was trained for 80 epochs with early stopping (patience=12) using the AdamW optimizer (initial learning rate=$3x10^{-4}$, weight decay=10^{-4}) and binary cross-entropy loss with label smoothing (epsilon=0.1). A OneCycleLR learning rate scheduler was employed with a peak learning rate of $1x10^{-3}$ at epoch 20. Training was conducted on four NVIDIA A100 GPUs in data-parallel mode using PyTorch DDP. The TCN and BiLSTM components were pre-trained for 20 epochs on the noise-free subset before full joint training. Gradient clipping (max norm=5.0) was applied to stabilise BiLSTM training. The final model was exported to TensorRT INT8 precision for edge deployment on the Jetson AGX Xavier.

D. Baseline Comparators

Four published anti-spoofing systems were re-implemented and trained on DefenceVoice-210K for fair comparison: (1) GMM-CQCC [6], (2) RawNet2 [8], (3) AASIST [17] — a state-of-the-art graph attention network, and (4) wav2vec-AASIST [10]. All baselines were trained using their published hyperparameters with the same data split as the proposed model, and evaluated with and without the ANSM pre-processing stage to isolate the contribution of noise suppression.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Detection Performance

Table I presents EER and AUC-ROC for all evaluated models on the 31,500-utterance test set under clean conditions and under battlefield noise (SNR=10 dB). The proposed ANRF-DVD achieves EER of 1.24% and AUC-ROC of 99.61% under clean conditions, and 2.37% EER / 98.83% AUC-ROC under 10 dB SNR — a degradation of only 1.13 percentage points, compared to 14.2 points for RawNet2 and 11.8 points for AASIST. Per-category analysis reveals that the hybrid TCN-BiLSTM-Transformer architecture achieves the highest gains over baselines for adversarially perturbed synthetic speech (C6: EER 1.91% vs. 8.44% AASIST), confirming the Transformer attention module's effectiveness at detecting subtle manipulated artefacts.

TABLE I. DEEPFAKE VOICE DETECTION PERFORMANCE COMPARISON

Model	EER (%) Clean	EER (%) SNR=10dB	AUC-ROC Clean	AUC-ROC SNR=10dB
GMM-CQCC [6]	6.81	24.37	96.12	81.43
RawNet2 [8]	4.12	18.29	97.84	87.61
AASIST [17]	2.74	14.51	98.76	90.23
wav2vec-AASIST [10]	2.31	11.08	99.04	92.87
Proposed ANRF-DVD	**1.24**	**2.37**	**99.61**	**98.83**

B. Inference Speed and Resource Utilisation

Table II summarises runtime performance metrics on the NVIDIA Jetson AGX Xavier (TensorRT INT8) and a reference x86 server (PyTorch FP32). The quantised ANRF-DVD achieves 27.4 ms per utterance on the Jetson Xavier, meeting the 30 ms maximum latency requirement for real-time voice authentication in tactical radio relay nodes. The

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405067

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

full-precision model achieves 11.2 ms on the GPU server. Power consumption at peak inference (6.8 W on Xavier) is compatible with deployments on battery-backed tactical communication terminals.

TABLE II. RUNTIME PERFORMANCE ON NVIDIA JETSON AGX XAVIER AND SERVER

Metric	ANRF-DVD (Jetson INT8)	ANRF-DVD (Server FP32)	RawNet2 (Jetson)	AASIST (Jetson)
Inference Time	27.4 ms	11.2 ms	18.6 ms	31.9 ms
Model Size	14.7 MB	58.2 MB	7.3 MB	22.1 MB
RAM Usage	1.2 GB	3.8 GB	0.6 GB	1.7 GB
Power Draw	6.8 W	42.3 W	4.1 W	9.2 W
EER (SNR=10dB)	2.37%	1.24%	18.29%	14.51%

C. Comparison with State-of-the-Art

Table III contextualises the proposed ANRF-DVD against published anti-spoofing systems. The proposed framework is the only system achieving sub-3% EER under noisy conditions while simultaneously supporting real-time edge deployment — a combination not achieved by any prior published work. The ANSM pre-processing contributes a mean absolute reduction of 9.4 percentage points in EER across all tested models under SNR=10 dB conditions, validating its generalisability as a noise-robustness augmentation module.

TABLE III. COMPARISON WITH STATE-OF-THE-ART DEEPFAKE VOICE DETECTION SYSTEMS

Study	EER(%)	Noise Robust	Real- Time	Edge Deploy	Dataset
Sahidullah et al. [6]	6.81	No	Yes	Partial	ASVspoof15
Tak et al. [8]	4.12	No	Yes	Partial	ASVspoof21
Jung et al. [9]	4.03	No	No	No	ASVspoof21
Wang et al. [10]	2.31	No	No	No	ASVspoof21
Khanjani et al. [14]	8.80	Partial	Yes	Yes	Custom
Proposed ANRF-DVD	**1.24**	**Yes**	**Yes**	**Yes**	**DefenceVoice-210K**

VI. DEPLOYMENT AND IMPACT ANALYSIS

A. Real-World Evaluation in Simulated Defence Scenarios

The ANRF-DVD was evaluated across three simulated tactical communication scenarios designed in collaboration with cyber-defence analysts: (S1) Base-Camp Secure Voice Network — static VoIP authentication with moderate background noise (SNR ~20 dB); (S2) Forward Operating Base HF Radio Relay — high channel distortion and compression artefacts (SNR ~10 dB); (S3) Urban Joint Operations Centre — mixed acoustic environment with multiple simultaneous speakers and variable SNR (5–25 dB). In each scenario, synthetic voice injection attacks were conducted using five state-of-the-art TTS and voice conversion systems not included in the training corpus, simulating zero-day adversarial conditions. Across all scenarios, the ANRF-DVD achieved a mean detection accuracy of 97.8%, with false alarm rate of 1.4% and missed detection rate of 2.2%. Scenario S2 presented the greatest challenge, with accuracy of 95.1%, attributed to severe codec artefacts that partially masked synthetic speech characteristics.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

B. Security and Operational Impact

Integration of the ANRF-DVD into the communications authentication pipeline is estimated to reduce successful voice spoofing intrusions by 94.3% compared to unaided human operator verification, based on a comparative red-team exercise conducted with 12 trained communications security analysts who achieved 61.2% detection accuracy against the same synthetic voice corpus. The framework's decision latency of 27.4 ms allows insertion into the authentication handshake without perceptible communication delay. Forensic logging via the DCAI module enables post-hoc attribution analysis, supporting intelligence and counter-intelligence workflows. The unit hardware cost (NVIDIA Jetson AGX Xavier + audio interface: approx. USD 899) is 73% lower than commercial biometric voice authentication appliances with equivalent throughput, providing a viable acquisition pathway for defence procurement.

C. Limitations and Future Work

The current framework exhibits the following limitations: (1) performance degrades under extreme channel noise below SNR=5 dB, which remains an open research problem; (2) the binary genuine/spoofed decision does not identify the specific spoofing algorithm, limiting forensic value; (3) the DefenceVoice-210K corpus, while comprehensive, does not include non-English language spoofing attacks relevant to multinational coalition operations. Future work will pursue multi-class spoofing algorithm attribution, cross-lingual deepfake detection using multilingual pre-trained acoustic models, and hardware-aware neural architecture search (NAS) to optimise the TCN-BiLSTM-Transformer pipeline for deployment on smaller embedded platforms (e.g., ARM Cortex-M55). Adversarial robustness against voice purification attacks and GAN-based anti-detection strategies will also be investigated. The planned integration of the ANRF-DVD with a broader multi-modal biometric authentication framework — combining voice, keystroke dynamics, and geolocation verification — is expected to further reduce authentication error rates to below 0.5%.

VII. CONCLUSION

This paper presented ANRF-DVD, an adaptive and noise-resilient AI framework for deepfake voice detection tailored to secure multi-agency defence communications. The proposed hybrid TCN-BiLSTM-Transformer architecture, combined with an adaptive noise suppression module and multi-domain acoustic feature fusion, achieved an EER of 1.24% under clean conditions and 2.37% under 10 dB SNR battlefield noise on the DefenceVoice-210K corpus — outperforming all compared state-of-the-art systems. A TensorRT-quantised deployment on the NVIDIA Jetson AGX Xavier demonstrated real-time inference at 27.4 ms per utterance with a power draw of 6.8 W, suitable for forward-deployed tactical communication nodes. Simulated defence scenario evaluations confirmed operational robustness at 97.8% detection accuracy against unseen synthetic voice attack systems. At a hardware cost of approximately USD 899 per node, the ANRF-DVD provides a cost-effective and scalable countermeasure to the growing threat of AI-generated voice spoofing in national defence and inter-agency coordination communications. The DefenceVoice-210K dataset and model weights will be released under restricted-access academic licence to support further research in secure voice communication.

VIII. ACKNOWLEDGMENT

The authors gratefully acknowledge the Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, for laboratory and computing facilities. The authors thank the volunteer military personnel who contributed speech recordings for the DefenceVoice-210K corpus under informed consent protocols. This research was supported in part by the Defence Research and Development Organisation (DRDO), Government of India, under Grant No. DRDO/CAIR/2024/AI-SEC/083, and by the Science and Engineering Research Board (SERB), Government of India, under Core Research Grant CRG/2023/004712.

REFERENCES

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016.
[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. IEEE ICASSP, 2018, pp. 4779–4783.
[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proc. ICML, 2021, pp. 5530–5540.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[4] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Beker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Tan, H. Zen, and Y. Wu, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Comput. Speech Lang., vol. 64, p. 101114, 2020.
[5] J. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements," in Proc. Odyssey, 2018, pp. 296–303.
[6] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," in Proc. Interspeech, 2015, pp. 2087–2091.
[7] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," in Proc. Interspeech, 2019, pp. 1013–1017.
[8] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in Proc. IEEE ICASSP, 2021, pp. 6369–6373.
[9] J. Jung, H.-S. Heo, H. Tak, H.-J. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in Proc. IEEE ICASSP, 2022, pp. 6367–6371.
[10] H. Wang, H. Dinkel, S. Wang, S. Kang, and Y. Qian, "Investigating self-supervised front-ends for speech spoofing countermeasures," in Proc. Odyssey, 2022, pp. 100–106.
[11] J. Yi, J. Fu, J. Tao, Z. Zheng, D. Zhang, C. Lv, and C. Fan, "Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms," in Proc. Interspeech, 2022, pp. 858–862.
[12] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep spectral features for anti-spoofing detection," in Proc. IEEE ICASSP, 2019,
pp. 6186–6190.
[13] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, T. Schneider, Z. B. Acar, T. Schneider, and N. Evans, "Preserving privacy with privacy-preserving speaker verification," IEEE Signal Process. Mag., vol. 36, no. 5, pp. 18–27, 2019.
[14] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," Front. Big Data, vol. 5, p. 1001063, 2023.
[15] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh, The Centre for Speech Technology Research, 2017.
[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in Proc. IEEE ICASSP, 2015, pp. 5206–5210.
[17] J. Jung, H.-S. Heo, H.-J. Shim, and H.-J. Yu, "Improved RawNet with feature map scaling for speaker verification," in Proc. Interspeech, 2020, pp. 3583–3587.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details