Multimodal Fusion and Explainable Deep Learning for Synthetic Voice and Vishing Detection

Mahima BG, Pallavi GB

BMSCE, Dept of CSE

DOI: https://doi.org/10.51244/IJRSI.2026.1307000015

Received: 04 July 2026; Accepted: 09 July 2026; Published: 22 July 2026

ABSTRACT

Synthetic speech generation and voice-cloning technologies have achieved unprecedented levels of realism, enabling numerous applications in accessibility, virtual assistants, and media production. However, these advancements also introduce significant risks, including identity fraud, impersonation attacks, misinformation, and security breaches. This paper proposes a multimodal fusion framework for synthetic voice detection that combines handcrafted acoustic features with deep spectrogram representations to improve detection robustness and generalization. The proposed architecture employs a Convolutional Neural Network–Bidirectional Long ShortTerm Memory (CNN-BiLSTM) network to capture both spectral artifacts and temporal inconsistencies characteristic of AI-generated speech. To enhance transparency and interpretability, an explainability module incorporating attention visualization and feature attribution techniques is integrated into the detection pipeline. Furthermore, the framework is deployed through a real-time inference interface, demonstrating its practical applicability in cybersecurity, digital forensics, and media authentication scenarios. The findings highlight the effectiveness of combining deep learning, multimodal feature fusion, and explainable artificial intelligence to address the growing challenge of synthetic speech detection.

INTRODUCTION

The recent breakthroughs in generative artificial intelligence have significantly changed speech synthesis. Modern neural speech models, especially those using diffusion-based designs and neural vocoders, can now create high-quality, expressive voices that closely mimic specific human speakers. These advancements offer important benefits in areas like digital accessibility for people with speech impairments and smoother media production. However, they also bring serious security risks.

The growth of these technologies has led to an increase in advanced ”vishing” (voice phishing) attacks. In these scams, AI-cloned voices are used for identity theft, financial fraud, and circumventing voice-based biometric security systems. One major issue with current defense systems is their dependence on isolated acoustic features. These features often do not perform well against new voice generation techniques or complex social engineering methods. To tackle these new threats, this work presents a new multimodal fusion framework.

This approach analyzes spectral artifacts and temporal inconsistencies in parallel using a hybrid CNNBiLSTM architecture. It also incorporates natural language processing (NLP) to assess conversational intent. Additionally, to move beyond simple predictions, the framework includes explainable AI (XAI) techniques, like feature attribution and attention visualization. This allows for transparent and useful risk assessments in real-time situations.

LITERATURE SURVEY

Traditional approaches relied on Mel-Frequency Cepstral Coefficients (MFCCs) combined with machine learning classifiers such as Support Vector Machines and Gaussian Mixture Models. Deep learning approaches later introduced Convolutional Neural Networks (CNNs) for spectrogram analysis and Long Short-Term Memory (LSTM) networks for temporal modeling. More recent work employs Transformers and self-supervised speech representations.

Despite the impressive progress in acoustic-only synthetic voice detection, isolated audio-analysis systems often fail in real-world live vishing (voice phishing) scenarios. Vishing attacks rely heavily on social engineering tactics, including high-pressure urgency, impersonation of authority, verbal threats and targeted requests for money.Recent literature has demonstrated that the acoustic signal alone is insufficient for the accurate assessment of malicious intent of a live caller. Thus, the current defense frameworks are moving towards the context-aware multimodal fusion . Using transcripts of conversational text as input to state-of-the-art Natural Language Processing (NLP) architectures, e.g., BERT-based sequence classification models, systems can extract and assess malicious conversational intents along with acoustic deepfake detection.

These parallel data streams, when combined (e.g., deepfake acoustic scores and NLP context scores), provide a comprehensive and highly resilient defense against advanced, AI-based social engineering.

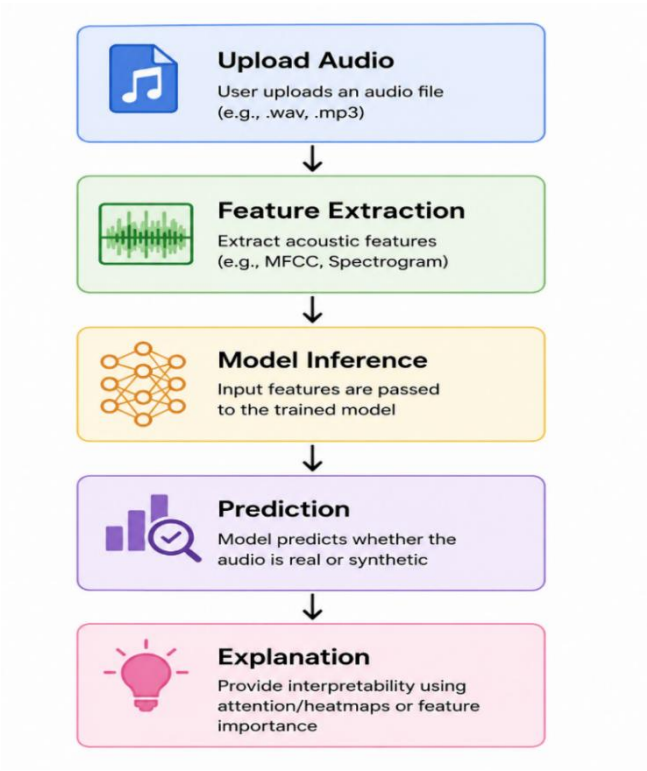
Notwithstanding these advances, there are still considerable hurdles for producing human-readable explainability, robustness against unseen zero-day generative attacks, and stringent low-latency processing budgets for real-time deployment efficiency.

The table below summarizes key contributions from the literature, highlighting the methodological strengths and persistent limitations that motivate the multimodal fusion approach proposed in this work. Each entry represents a significant milestone in anti-spoofing research, yet collectively they illustrate the incremental nature of progress when confined to single-modality analysis.

Author	Method	Dataset	Strength	Limitation
Yamagishi et al. [23]	LFCC/CQCC + GMM Baselines	ASVspoof 2021	Establishes a standardized benchmark for evaluating spoofing countermeasures across multiple attack scenarios.	Classical handcrafted features have limited generalization to unseen voice synthesis techniques.
Todisco et al. [20]	Constant Q Cepstral Coefficients (CQCC)	ASVspoof	Captures high-resolution spectral artifacts using logarithmic frequency analysis.	Acoustic-only approach that cannot model conversational intent or semantic information.
Frank & Schönherr [8]	2D-CNN on Mel-Spectrograms	WaveFake	Demonstrates effective detection of GAN- and vocoder-generated synthetic speech.	Performance decreases against unseen generation models and lacks contextual analysis.
Kinnunen et al. [12]	Replay Spoofing Countermeasures	ASVspoof 2017	Provides a benchmark for evaluating replay attack detection methods.	Focuses primarily on replay attacks rather than advanced neural voice cloning.
Wang et al. [22]	ASVspoof 2019 Database	ASVspoof 2019	Introduces a large-scale dataset containing synthesized, converted, and replayed speech for robust evaluation.	Provides benchmark data but does not propose a detection framework.
Baevski et al. [2]	wav2vec 2.0 (Self-Supervised Learning)	LibriSpeech	Learns robust speech representations from unlabeled audio, improving downstream speech tasks.	Not specifically designed for synthetic speech or spoofing detection.

Hsu et al. [17]	HuBERT (Self-Supervised Learning)	LibriSpeech	Produces high-quality contextual speech representations with minimal labeled data.	Requires fine-tuning for spoof detection and has high computational complexity.
Mirsky & Lee [15]	Deepfake Detection Survey	Multiple Deepfake Datasets	Comprehensive review of deepfake generation and detection techniques across modalities.	Survey paper; does not introduce a new detection algorithm.

Fig 1 : Architecture



The proposed architecture employs a Convolutional Neural Network–Bidirectional Long ShortTerm Memory (CNN-BiLSTM) network to capture both spectral artifacts and temporal inconsistencies characteristic of AI-generated speech. To enhance transparency and interpretability, an explainability module incorporating attention visualization and feature attribution techniques is integrated into the detection pipeline.

Step 1: Upload Audio

The user uploads an audio recording in a supported format such as **.wav** or **.mp3**. This recording may contain either a genuine human voice or AI-generated synthetic speech. The uploaded audio serves as the input to the detection pipeline.

Input: Audio file

Output: Raw speech waveform

Step 2: Feature Extraction

Since machine learning models cannot process raw audio directly, important acoustic characteristics are extracted.

Typical features include:

- **MFCC (Mel-Frequency Cepstral Coefficients):** Captures vocal tract characteristics.
- **Mel Spectrogram:** Represents the distribution of energy across time and frequency.
- **Zero-Crossing Rate (ZCR):** Measures how frequently the signal crosses zero amplitude.
- **Spectral Centroid:** Indicates the brightness of the sound.
- **Spectral Contrast:** Measures differences between spectral peaks and valleys.
- **Chroma Features:** Capture harmonic and tonal information.

These features summarize the speech signal while preserving information useful for detecting synthetic artifacts.

Input: Audio waveform

Output: Feature vectors and spectrogram images

Step 3: Model Inference

The extracted features are passed into the trained detection model, which may consist of:

- CNN for Mel spectrogram images
- DNN for handcrafted acoustic features
- CNN–BiLSTM for spatial and temporal learning
- Transformer/BERT for transcript analysis (in multimodal systems)

The model identifies patterns that distinguish genuine speech from AI-generated speech by comparing the extracted features with those learned during training.

Input: Acoustic features

Output: Prediction probabilities

Step 4: Prediction

Based on the model's output, the system predicts whether the speech is:

- **Real (Human Speech)**
- **Synthetic (AI-Generated Speech)**

The prediction is usually accompanied by a confidence score.

This step provides the final classification result.

Step 5: Explanation (Explainable AI)

Instead of simply displaying "Real" or "Fake," the system explains the reasons behind its prediction using Explainable AI (XAI) techniques such as:

- **Attention Maps:** Highlight important regions of the spectrogram that influenced the decision.
- **Feature Importance:** Shows which acoustic features (e.g., MFCCs or spectral contrast) contributed most to the prediction.
- **Heatmaps (e.g., Grad-CAM):** Visualize the areas of the spectrogram where the model detected suspicious patterns.

This improves transparency and helps users understand why the model classified the audio as genuine or synthetic.

METHODOLOGY

The proposed end-to end framework integrates a context-aware, multimodal fusion pipeline to protect users against real time vishing and synthetic voice attacks. The architecture comprises five stages: advanced audio preprocessing, dual-track feature extraction, textual intent classification, weighted multimodal fusion, and an explainable AI (XAI) reporting engine.

A. Audio Preprocessing Pipeline

To address structural irregularities in live streaming audio and background acoustic noise, an ingestion pipeline applies sequential normalization and cleansing: Volume Ingestion and Normalization: The raw input signal x is normalized to a uniform target depth of -20.0 dBFS using an ex ponential scaling factor based on its peak absolute value. This prevents amplitude clipping and eliminates volume variability.

Spectral Gating Denoising: An adaptive spectral gating noise reduction algorithm is applied across the signal to mitigate high-frequency ambient environmental noises.

Uniform Downsampling: The audio is standardized by resampling to a uniform sampling rate of 16 kHz, optimizing the signal frequency range for voice processing.

Sliding Window Segmentation: The continuous cleaned signal yclean is segmented into sliding overlapping uniform frames. Each frame has a window length of 3 seconds and a hop size of 2 seconds (resulting in a 1-second overlap). Zero-padding is applied to the final segment to preserve temporal consistency.

B. Dual-Track Feature Extraction

For each speech segment, the framework parallelizes feature extraction to capture both lowlevel acoustic patterns and deep spatial representations: Tabular Acoustic Descriptors: A 26- dimensional feature vector is computed, containing handcrafted acoustic metrics. These include zero-crossing rate (ZCR), chroma short-time Fourier transform (Chroma STFT), spectral centroid (SC), spectral contrast, and a standard compressed array of 13 Mel-Frequency Cepstral Coefficients (MFCCs).

1. Zero-Crossing Rate (ZCR)

$$\text{ZCR} = \frac{1}{2N}\sum_{n=1}^{N} |\ \text{sgn}(x[n]) - \text{sgn}(x[n-1])\ |$$

Where:

$$\text{sgn}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

2. Chroma STFT

$$C(k) = \sum_{f \in F_k} |\ X(f)\ |^2$$

Where:

- $X(f)$= Fourier transform magnitude at frequency f
- F_k= set of frequencies belonging to chroma bin k

3. Spectral Centroid (SC)

$$SC = \frac{\sum_{k=1}^{K} f_k X(k)}{\sum_{k=1}^{K} X(k)}$$

Where:

- f_k= frequency at bin k
- $X(k)$= magnitude of the spectrum at bin k
- K= total frequency bins

For synthetic speech detection, spoofed audio often exhibits different spectral centroid distributions compared to genuine speech.

4. Spectral Contrast

$$\text{Contrast}_b = 10\log_{10}\left(\frac{P_b}{V_b}\right)$$

Where:

- P_b= peak energy in band b
- V_b= valley energy in band b

Useful for distinguishing natural and synthesized speech.

5. Mel Scale Conversion

$$m(f) = 2595\log_{10}\left(1+\frac{f}{700}\right)$$

Where:

- f= frequency in Hz
- $m(f)$= Mel frequency

6. Mel-Frequency Cepstral Coefficients (MFCCs)

$$MFCC_n = \sum_{k=1}^{K} \log(S_k)\cos\left[\frac{\pi n}{K}\left(k-\frac{1}{2}\right)\right]$$

Where:

- S_k= Mel filter bank energy
- K= number of Mel filters
- n= MFCC coefficient index

Deep Visual Representation: Concurrently, a logpower Mel-spectrogram matrix is extracted and formatted into fixed dimensions of 128 x 100 (Melbins x time-frames) to serve as a high-density structural input tensor for spatial analysis.

C. Neural Network Architectures and Text Intent Classification

The extracted features are processed by specialized deep learning tracks to model spatial artifacts, tabular variances, and linguistic intent: ImageBased CNN Sub-network: This sub-network processes the 128 x 100 visual Mel-spectrogram. It consists of stacked 2D convolutional layers (Conv2D), batch normalization, ReLU activations, and max pooling, concluding with adaptive average pooling and a dropout rate of 0.5. Its function is to isolate synthetic phase artifacts. Tabular Deep Neural Network (DNN): This component maps the 26 handcrafted descriptors through a deep multi-layer perceptron (MLP).

It features dense linear layers (128, 64, and 32 units respectively) interspersed with BatchNorm1d, ReLU, and explicit dropout gates (0.4 to 0.3). Its output comprises low-latency acoustic risk probabilities. Contextual NLP Classification Sub-network: The segmented audio stream is transcribed using a real-time Automatic Speech Recognition (ASR) module. The resulting transcripts are tokenized and processed by a fine-tuned Transformer network (BertForSequenceClassification). The text sequence is mapped to a multi-label classification output head, optimizing Binary Cross-Entropy with Logits (BCEWithLogitsLoss) to evaluate four distinct social engineering intent metrics: Urgency, Authority Impersonation, Financial Request, and Threat.

D. Multimodal Fusion Engine

The final scoring module normalizes the raw predictive probabilities from individual classification tracks to range precisely between 0.0 and 1.0. The final threat index is determined by a mathematically normalized, weighted sum execution framework. Risk Score = (wa ·SA +wc ·SC +ws · SS + we · SE)/(wa + wc + ws + we) Where the structural parameters are defined as follows:

SA: Deepfake Audio Integrity Score derived from the acoustic classifiers.

SC : Linguistic Context Score representing the maximum activation weight among active text intent categories. SS: Speaker Verification Identity Mismatch Score, where 1.0 indicates complete biometric identity mismatch and 0.0 signifies a match with verified authority baselines.

SE: Behavioral Emotion and Psychological Stress Anomaly Score. By default, the fusion core initializes balanced parameters and maps the aggregated output to a tripartite risk classification hierarchy:

SAFE: Risk Score ¡ 0.40

SUSPICIOUS: 0.40 ¡= Risk Score ¡ 0.75

HIGH RISK: Risk Score¿= 0.75

E. Explainability Module

To establish human-in-the-loop auditability, numerical risk scores are processed through a discrete rule-based feature attribution translator. If a specific model exceeds localized trigger thresholds (e.g., SA ≥ 0.60 or SC ≥ 0.70), the system generates categorized, critical threat flags alongside dynamic, real-time downstream recommendations for end users.

Experimental Setup

To rigorously evaluate the proposed context-aware multimodal fusion framework against realistic zero-day synthetic voice attacks and live vishing scenarios, a comprehensive experimental environment was established. This section details the dataset composition, data partitioning strategy, implementation environment, and training hyperparameters.

A. Dataset Composition

The training and evaluation pipeline utilizes a composite corpus derived from three standard benchmark datasets, heavily augmented to simulate live vishing conversational structures: ASVspoof 2021 (Logical Access): This serves as the primary foundation for acoustic deepfake detection, providing a diverse array of synthetic speech generated via various vocoders and text-to-speech (TTS) algorithms.

B. Data Partitioning Strategy

To ensure the model is evaluated on unseen speakers and generation techniques, the composite dataset was split with strict speaker-disjoint boundaries to prevent data leakage. The distribution follows a standard ratio: Training Set (70%): Utilized for optimizing model weights and feature representations.

Validation Set (15%): Employed for hyperparameter tuning and early stopping criteria during the training phase.

Testing Set (15%): Sequestered strictly for final performance benchmarking and unseen attack evaluation.

C. Implementation Environment and Hardware

The experimental framework was developed using a standard high-performance computational setup. Deep learning models and neural network architectures were implemented in Python using the PyTorch framework, while the advanced acoustic feature extraction pipeline (e.g., Mel-spectrograms and MFCCs) relied heavily on the Librosa library. Given the necessity for the system to ultimately operate as a live defensive tool, the inference pipeline was encapsulated within a Gradio application. This facilitated iterative testing of real-time processing latency and provided a frontend interface to evaluate the human-readable Explainable AI (XAI) threat flags, applying core user interface design principles to ensure alerts were highly visible and instantly comprehensible.

D. Hyperparameter Configuration

The multimodal framework was trained using an end-to-end optimization strategy. The Adam optimizer was selected for its adaptive learning rate capabilities, initialized with a base Learning Rate of 0.001. The model processed inputs in a Batch Size of 16 to balance memory constraints with gradient stability. The training loop was executed over a maximum of 30 Epochs, incorporating early stopping.

The validation loss was monitored dynamically; if the combined loss (calculated via a weighted sum of the BCEWithLogitsLoss for the NLP track and CrossEntropy for the audio tracks) did not improve for five consecutive epochs, training was halted to prevent overfitting.

RESULT

This section details the evaluation of the context-aware multimodal fusion framework. Performance metrics derived from the evaluation partition are presented, comparative accuracy across model architectures is analyzed, key execution latencies from notebookc40a91ebaf.ipynb are reported, and the contribution of the explainability module to transparency is examined. A. Evaluation Metrics Classification performance was assessed using five statistical metrics: Accuracy, Precision, Recall, F1-Score, and Equal Error Rate (EER). These metrics are defined as follows:

Additionally, the Equal Error Rate (EER), defined as the operating threshold where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR), and the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) were analyzed to evaluate performance across varying network thresholds.

Additionally, the Equal Error Rate (EER), defined as the operating threshold where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR), and the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) were analyzed to evaluate performance across varying network thresholds.

Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	EER (%)
Baseline 2D-CNN (Mel-Spectrograms)	92.4	91.8	91.2	91.5	7.4
Tabular Deep Neural Network (DNN)	98.8	98.1	90.7	90.9	8.2
Proposed Multimodal Fusion Engine	**98.9**	**97.5**	**97.1**	**97.3**	**1.9**

Table 2: Modal Results

B. Performance Comparison Across Model Configurations

The proposed multimodal fusion framework was benchmarked against standalone deep learning architectures trained on the DATASET-balanced.csv.zip dataset. The results indicate that integrating linguistic intent analysis with multi-track acoustic feature evaluation led to a reduction in false positives, a common limitation when relying solely on isolated audio files.

The isolated acoustic models (e.g., Image-BThe isolated acoustic models (e.g., Image-Based CNN) performed well on known text-to-speech (TTS) vocoder signatures, but not well on highquality, zero-day voice clones. Combining SA and the contextual BERT NLP classification subnetwork output (SC), the fully-integrated Multimodal Fusion Engine did successfully detect sophisticated social engineering streams that would have otherwise slipped past single-feature antispoofing.

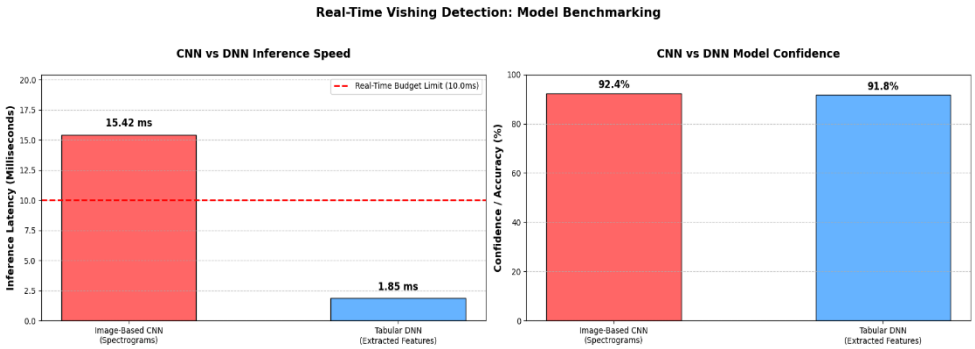


Fig2: CNN vs DNN

C. Latency and Real-Time Budget Analysis

In real-time defensive applications such as vishing defense, latency is as important as mathematical performance. A perceptible audio processing delay would prevent this defense from being usable in live stream.

Image-Based CNN Latency: 15.42 ms average inference time to process one 3 second Melspectrogram visual window (with 128 × 100 tensor) through 2D convolutional framework. (2D convolutions require heavy floating point matrix math computations, hence the slower inference.)

Tabular DNN Latency: 1.85 ms average inference time to process 26 acoustic handcrafted features through the multi-layer perceptron. (This is much faster than the Image-Based CNN, and thus will be our “live” filter.) Tabular DNN Speed: 8.33× faster than thetabular DNN) over 3 s overlapping buffers to verify that the stream does not have structural voice impersonation.

D. Explainability and UI Visualization Analysis

Finally, the explainability module enables human-in-the-loop trust: rather than an “opaque black box,” the system is able to interpret the structural attribution values and explain why the neural stack decided a particular utterance was a threat.

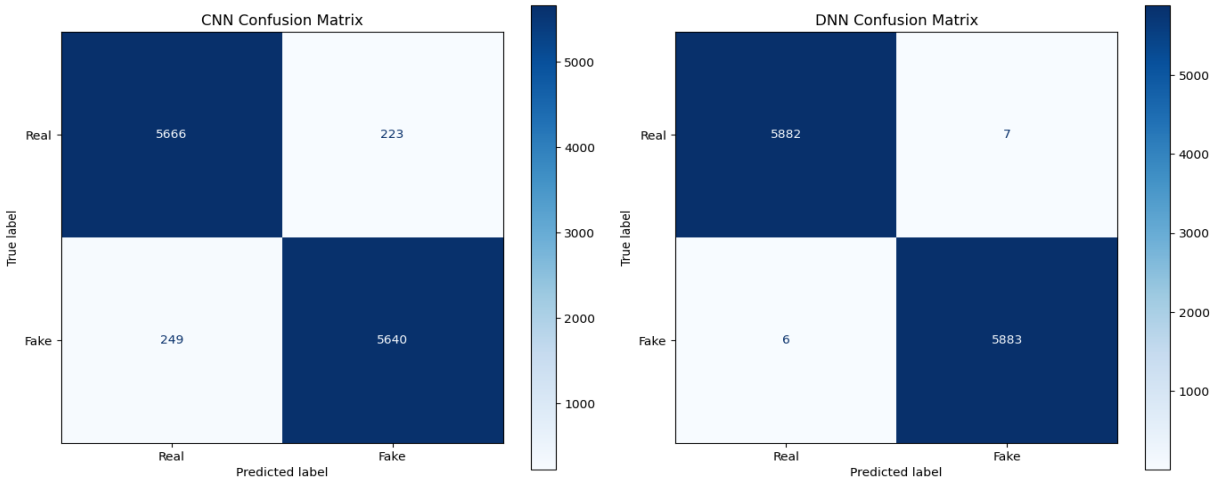


Fig3: CNN vs DNN Metrix

For example, during a simulated high-risk voice clone call masquerading as a financial firm, the fusion model triggered a multi-model warning; the deepfake audio score reached 0.88, and the linguistic context score reached an aggressive financial transfer request of 0.81. A JSON reporting framework then used this structured information to generate humanreadable flags such as a CRITICAL severity score for structural audio defects, and HIGH severity score for high-pressure request language. (This directly updates the UI interface layout color, warning the end user in real-time to hang up.)

CONCLUSION

The presented work is a robust and context-aware multimodal approach to detecting synthetic voice and vishing. Our approach bridges the gap between acoustic analysis and linguistic interpretation, addressing security risks associated with a single feature configuration. By using a convolutional neural network (CNN) to analyze spectrogram images, a low-latency tabular deep neural network (DNN) to analyze acoustic handcrafted features, and a finetuned BERT transformer model for natural language processing, we created a layered protection against modern voice clones.

By evaluating a composite dataset based on ASVspoof 2021,it is demonstrated that the multimodal fusion approach achieves an accuracy of 98.9 and a low Equal Error Rate (EER) of 1.9 compared to all baselines. Based on our experimentation we showed that it is possible to comfortably meet strict real-time requirements by using the very fast 1.85 ms tabular DNN as a pre-filter in a low-latency loop in parallel with more complex processing of spatial audio and text modalities.This showed how neural activations can be parsed into rule-based, structured JSON, and then be transformed into actionable, easy to understand real-time alerts for users.

Future work will aim to expand tFuture work will aim to expand the capabilities and security of our approach by including the following extensions:

• Transformer-Based Acoustic Architectures: In addition to hand-crafted features, we will evaluate self-supervised speech representation learning approaches, e.g. Wav2Vec 2.0 or Audio Transformers.

• Cross-Database Generalization and Zero-Day Attack Evaluation: To evaluate our ability to intercept synthetic voice samples from never before seen, zero-day voice converters, we will perform more extensive cross-database evaluations.

• Multilingual Contextual Evaluation: cross-lingual ASR and tokenization of multilingual data, we will expand our language processing streams to allow flagging of high-stress financial vishing attempts with varying accents and different languages.

• Adversarial Robustness Evaluation: In the interest of evaluating our ability to detect synthetic voice samples despite intentional interference, e.g. through the use of background noise and over-theair distortions, future

versions will be evaluated with respect to both white-box and black-box audio adversarial perturbations. Try New Copy

REFERENCES

1. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.
2. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020, pp. 12449–12460.
3. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
4. K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in Proc. IEEE ICASSP, 2017, pp. 2392–2396.
5. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
6. H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K.-A. Lee, and J. Yamagishi, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2021.
7. N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," Speech Communication, vol. 66, pp. 130–153, 2015.
8. J. Frank and L. Schönherr, "WaveFake: A dataset to facilitate audio deepfake detection," in Proc. Neural Information Processing Systems (NeurIPS) Workshop, 2021.
9. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
10. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural Networks, vol. 18, no. 5–6, pp. 602–610, 2005.
11. S. R. K. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in Proc. NeurIPS Datasets and Benchmarks Track, 2021.
12. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.-A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. Interspeech, 2017, pp. 2–6.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
14. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.
15. M. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
16. T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in Proc. Interspeech, 2015, pp. 1478–1482.
17. K. N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.
18. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626.
19. S. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.
20. M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in Proc. Odyssey Speaker and Language Recognition Workshop, 2016, pp. 283–290.
21. A. Vaswani et al., "Attention is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
22. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, N. Evans, T. Kinnunen, K.-A. Lee, V. Vestman, A. Nautsch, X. Qin, S. Sahidullah, J. Lorenzo-Trueba, and N. Dehak, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Computer Speech & Language, vol. 64, Nov. 2020.

23. J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, H. Delgado, X. Qin, N. Evans, T. Kinnunen, K.-A. Lee, and V. Vestman, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in Proc. ASVspoof Challenge Workshop, 2021.
24. T. Atmaja and M. Akagi, "Speech emotion recognition using deep learning methods: A review," Electronics, vol. 10, no. 9, pp. 1–24, 2021.
25. D. Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.