International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Deepfake Voice Based Scam Detection in Banking Calls Using AI and ML

Dr.K.Karuppasamy
Head of the Department
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
karuppusamyrvs@gmail.com

R.Thenmalar
Assistant Professor
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
thenmalarcbe@gmail.com

S. Jayaraman
712822104015
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
Jayaraman14122004@gmail.com

P. Gokul
712822104010
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
Gokulp102005@gmail.com

R. Janarthan
712822104014
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
janaallan6317@gmail.com

M. Ajay Makeshwaran
712822104703
Dept of Computer Science & Engineering
RVS College of Engineering & Technology,
Coimbatore, India.
ajaymakeshwaran@gmail.com

Abstract:

With the recent developments in Artificial Intelligence (AI), realistic synthetic speech can now be produced through voice cloning and neural network-based text-to-speech systems. However, the security threats associated with such technologies are critical in the banking and financial sector. Deepfake voice-based scam calls are a recent form of sophisticated fraud in which attackers pose as bank officials and extract critical information from the victim.

The current paper presents a real-time deepfake voice detection system for the banking call scenario. This system can stream audio signals in real-time and use a hybrid Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) model to classify the audio signals as real or fake. This architecture can also generate alerts to prevent financial fraud. Experimental results on benchmark datasets have proven the effectiveness of the proposed system.

The proposed system can greatly enhance communication security and can be used as a robust defense mechanism against AI-based voice cloning technologies scams

1. Introduction

Recently, Artificial Intelligence has transformed the field of speech synthesis technology. Advanced neural networks can produce human speech that closely resembles tone, pitch, accent, and speech patterns of humans. This has greatly impacted various applications, including virtual assistants, accessibility tools, and customer support systems. However, the misuse of speech synthesis technology has posed a major cybersecurity threat.

Deepfake voice technology helps attackers to create the clone of a particular person's voice with fewer audio samples. In the banking sector, attackers use this technique to trick victims, who are bank officials or trusted contacts, to reveal sensitive information such as One-Time Passwords (OTPs), account credentials, and transaction information. Deepfake voice scams are not easy to detect, unlike traditional scam calls.

Currently, fraud detection systems used in the banking sector focus on transaction monitoring and behavioral analytics. These systems, although efficient in fraud detection, do not verify the authenticity of voice communication in a call. This creates a major loophole in voice-based customer interactions.

In order to mitigate this problem, this paper proposes a real-time AI-based deep fake voice detection system for banking calls. In the proposed system, the voice detection between human and AI-generated voice will be carried out based on the analysis of acoustic and temporal features of the voice. In the proposed system,

the deep learning model will be implemented for the detection of voice.

The main contributions of this work are:

1. Design of a real-time deepfake detection architecture for banking calls.
2. Implementation of a hybrid CNN-RNN model for robust speech classification.
3. Integration of an alert system for immediate fraud prevention.
4. Evaluation using benchmark spoofing datasets.

2. RELATED WORKS

The rapid development of speech synthesis techniques has led to more research in the detection of synthetic and manipulated audio signals. Deepfake voice detection, which is also known as spoofed speech detection and presentation attack detection, is considered a key area of research in speech processing and cybersecurity.

The traditional methods of synthetic speech detection were based on statistical models and signal processing. In particular, the traditional methods used Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) to analyze the distortions in the frequency spectrum of the spoofed audio. Nevertheless, these methods were limited in terms of generalization, especially with the use of advanced neural-based voice cloning techniques.

The introduction of deep learning also improved the detection performance. Wu et al. have used deep neural network (DNN) classifiers based on spectral features such as Mel frequency cepstral coefficients (MFCCs) and log power spectra. Their research proved that deep learning outperforms traditional statistical approaches in synthetic speech pattern detection.

One of the major milestones in this field was the ASVspoof Challenge proposed by Todisco et al. This challenge was used to obtain benchmarking datasets for evaluating spoofing countermeasure systems. The challenge was comprised of various types of attacks, including text-to-speech (TTS) attack, voice conversion (VC) attack, and replay attack. The ASVspoof 2019 and ASVspoof 2021 datasets proposed logical access (LA) and physical access (PA) types of attacks.

In addition to magnitude-based spectral features, Lavrentyeva et al. have also worked with phase-based features. Their findings have demonstrated the importance of phase-based features in speaker verification systems, where synthetic speech may not accurately represent the phase information. This ensures the effectiveness of phase-based features in spoofing detection.

Snyder et al. have proposed x-vector-based embeddings for speaker verification systems. This concept was further applied to spoofing detection systems. In this approach, a speaker representation is created using embeddings and classified using backend classifiers like SVM.

Recently, convolutional neural networks (CNN) have been used extensively for spectrogram-based classification approaches. In CNN models, speech signals are represented as images, and spatial representations of speech signals are learned. However, it is difficult to fully represent temporal dependencies in speech signals using CNN models, as speech signals are sequential in nature.

Recently, researchers proposed a new approach to overcome the limitations of CNN models in speech signal processing, in which a new type of neural network, Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) networks, is combined with CNN models to improve the accuracy of speech signal detection, as shown in the study by Zhang et al.

The other emerging trend is the use of transformer models and other self-supervised learning techniques like wav2vec. These models are effective in learning speech representations from large speech datasets and have shown good performance in spoofing detection. Nevertheless, the computational cost of these models might affect their use in real-time banking environments.

In the context of financial fraud detection, Singh and Patel highlighted the significance of AI-based monitoring systems to combat cybercrime in banking communication. Although there is established research in financial fraud detection systems, little attention has been given to real-time voice authentication in communication.

Despite significant progress, several limitations remain in existing works:

1. Many systems operate offline and are not optimized for real-time deployment.
2. Some models require high computational resources unsuitable for banking call centers.
3. Limited integration with alert mechanisms for immediate fraud prevention.
4. Insufficient focus on banking-specific threat scenarios.

The proposed system seeks to eliminate the limitations of the existing approaches through the integration of a lightweight hybrid deep learning architecture, audio processing, and instant alert generation. Unlike other spoof detection approaches, the proposed system has been designed for the banking call environment, ensuring real-time fraud mitigation.

3.THE PROPOSED CONSTRUCTION

III. Proposed Construction

The proposed deep fake voice detection system is designed to be a real-time, modular, and scalable framework that is particularly tailored for the banking call environment. The architecture of the system incorporates speech processing and hybrid deep learning models to identify AI-generated synthetic voices while communicating in real-time. The construction of the system is layered.The system consists of six major components:

1. Audio Acquisition Module
2. Preprocessing Module
3. Feature Extraction Module
4. Deep Learning Classification Module
5. Decision and Alert Module
6. Data Storage and Administrative Module

Each module performs a specific function within the detection pipeline.

A. Audio Acquisition Module

The first stage of the system receives the incoming and outgoing banking calls via a secure communication interface. Here, audio signals are recorded in real time and converted to a standard format such as WAV with a sampling rate of 16 kHz and 16-bit resolution.

For privacy and compliance purposes, the audio signals are encrypted during transmission and temporarily buffered for processing. This minimizes privacy risks since audio signals are not permanently stored unless deemed suspicious.

B. Preprocessing Module

Telephonic speech signals often contain background noise, echo, and channel distortions. Therefore, preprocessing is applied to enhance signal quality before feature extraction.

The preprocessing stage includes:

1. Noise Reduction: Spectral subtraction and filtering techniques are applied to remove background noise.
2. Silence Removal: Non-speech segments are eliminated to focus analysis on active voice frames.
3. Normalization: Amplitude normalization ensures uniform loudness levels across different calls.
4. Framing and Windowing: The signal is divided into short overlapping frames (e.g., 25 ms with 10 ms overlap) and windowed using Hamming window to reduce spectral leakage.

These steps improve model robustness and reduce false detections.

C. Feature Extraction Module

Feature extraction plays a critical role in distinguishing human speech from synthetic speech. Deepfake voices often exhibit subtle spectral and temporal inconsistencies due to neural generation artifacts.

The following acoustic features are extracted:

- Mel-Frequency Cepstral Coefficients (MFCCs)
- Spectral Centroid and Bandwidth
- Zero-Crossing Rate (ZCR)
- Pitch and Energy Contours
- Phase-Based Features

where each $f_i$corresponds to a specific acoustic parameter. These features are combined to form a high-dimensional representation of the speech sample.

The extracted feature matrix is then converted into spectrogram representations for CNN processing.

D. Deep Learning Classification Module

The classification engine is built using a hybrid CNN-RNN architecture to capture both spatial and temporal speech characteristics.

1) Convolutional Neural Network (CNN)

The CNN processes spectrogram images and extracts local frequency patterns. Convolutional layers identify texture differences between natural and synthetic voices, while pooling layers reduce dimensionality.

2) Recurrent Neural Network (RNN-LSTM)

Speech signals are sequential in nature. The LSTM layer captures long-term temporal dependencies, such as unnatural rhythm or inconsistent speech dynamics found in AI-generated voices.

3) Fully Connected Layer and Softmax

The final classification is performed using a dense layer with Softmax activation:

The model is trained using cross-entropy loss and optimized using Adam optimizer. Training is performed using benchmark datasets such as ASVspoof 2019 and 2021 to ensure generalization.

E. Decision and Alert Module

After classification, the predicted probability is compared against a predefined threshold T.

When synthetic speech is detected, the system performs:

- Immediate user notification
- Alert display to banking staff
- Logging of suspicious call
- Optional automatic call termination

This real-time response prevents customers from disclosing sensitive information.

F. Data Storage and Administrative Module

The system maintains a secure database to store:

- Detection logs
- Feature statistics
- Alert records
- System performance metrics

A centralized Admin Dashboard provides:

- Monitoring of flagged calls
- Threshold adjustment
- Model performance visualization
- System configuration management

Role-based access control ensures that only authorized personnel can access sensitive logs.

G. System Workflow Summary

The overall workflow of the proposed construction can be summarized as:

1. Capture live banking call audio
2. Preprocess signal
3. Extract discriminative features
4. Classify using hybrid CNN-RNN model
5. Trigger alert if synthetic speech detected
6. Store detection results for audit and monitoring

This modular construction ensures scalability, adaptability, and ease of integration with existing banking communication systems.

4.System Model

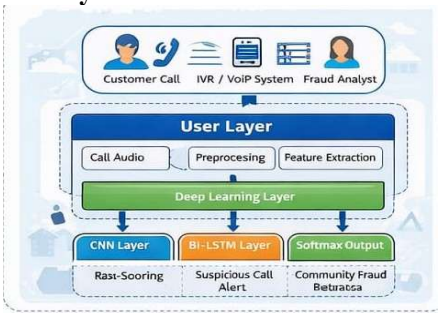
The system model defines the structural architecture and interaction between different components of the proposed deepfake voice detection framework. The objective of the model is to ensure secure, real-time detection of AI-generated voices in banking calls while maintaining scalability and integration capability with existing banking infrastructure.

The system follows a layered and modular architecture consisting of five primary layers:

1. User Layer
2. Audio Processing Layer
3. Machine Learning Layer
4. Data Management Layer
5. Alert and Integration Layer

Each layer performs a specific function and interacts systematically to ensure efficient detection and response.

A. User Layer



The User Layer represents all entities interacting with the system. It includes:

- Banking customers
- Call center representatives
- System administrators

Customers and bank officials participate in voice communication, while administrators monitor system performance through the dashboard interface. This layer serves as the input and output boundary of the system.

B. Audio Processing Layer

The Audio Processing Layer is responsible for capturing and preparing speech signals for analysis. It consists of:

1. **Audio Capture Module** – Records live call audio streams securely.
2. **Preprocessing Module** – Performs noise reduction, silence removal, normalization, and segmentation.
3. **Feature Extraction Module** – Extracts acoustic features such as MFCCs, spectral features, pitch, and phase information.

where F is the extracted feature vector and E denotes the extraction process.

This layer ensures that only meaningful and enhanced speech features are forwarded to the classification module.

C. Machine Learning Layer

The Machine Learning Layer forms the core intelligence of the system. It implements a hybrid CNN-RNN architecture to classify speech signals as either genuine human speech or AI-generated synthetic speech.

The CNN component extracts spatial representations from spectrogram images, while the RNN (LSTM) captures sequential and temporal speech characteristics. The final output is obtained using a Softmax function that provides classification probabilities.

This layered learning approach improves detection accuracy by combining spectral and temporal information.

D. Data Management Layer

The Data Management Layer handles storage, logging, and system monitoring. It includes:

- Detection logs database
- Suspicious call records
- Model performance metrics
- User access control records

A relational database such as MySQL or PostgreSQL is used for structured data storage. Sensitive information is encrypted to ensure compliance with security standards.

This layer ensures traceability and supports future auditing and system improvements.

E. Alert and Integration Layer

The Alert and Integration Layer enables real-time response and system interoperability. When synthetic speech is detected, the system performs:

- Immediate user notification
- Warning alert to banking staff
- Optional automated call termination
- Logging of incident

Additionally, REST APIs allow integration with:

- Banking server systems
- Mobile banking applications
- Telecom monitoring systems

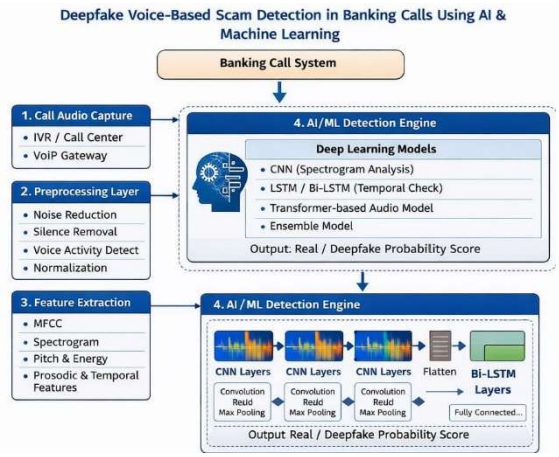
This ensures seamless deployment within existing financial infrastructures.

F. Overall System Interaction

The overall interaction of the system can be summarized as:

User Call → Audio Capture → Preprocessing → Feature Extraction → CNN-RNN Classification → Decision Logic → Alert Generation → Data Storage

The modular nature of the system model allows independent optimization of each layer. It also supports scalability for high-volume call environments such as large banking call centers.



5. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to **Dr. K. Karuppasamy**, Head of the Department of Computer Science and Engineering, RVS College of Engineering and Technology, Coimbatore, for his continuous encouragement, valuable guidance, and constant support throughout the development of this project. His leadership and insights greatly contributed to the successful completion of this research work.

The authors also extend their heartfelt thanks to **Ms. R. Thenmalar**, Assistant Professor, Department of Computer Science and Engineering, for her constructive feedback, technical suggestions, and consistent mentoring during every phase of the project. Her expertise in machine learning and system design significantly improved the quality and clarity of this work.

We sincerely acknowledge the support provided by the faculty members of the Department of Computer Science and Engineering for their cooperation and encouragement. The infrastructure and laboratory facilities provided by RVS College of Engineering and Technology played a crucial role in carrying out experiments and model evaluations effectively.

The authors would also like to acknowledge the researchers and organizers of the ASVspoof Challenge for providing benchmark datasets that supported the training and validation of the proposed deepfake detection model.

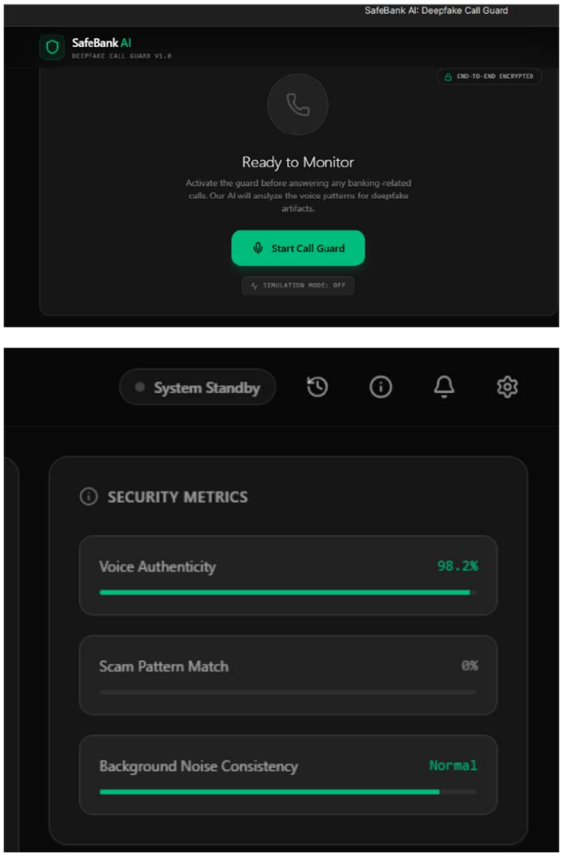
Finally, we express our gratitude to our friends and family members for their moral support and motivation, which helped us complete this research successfully.

6.CONCLUSION

This paper presented a real-time deepfake voice detection system designed for banking call environments. The proposed framework integrates speech preprocessing, acoustic feature extraction, and a hybrid CNN-RNN model to classify speech as genuine or AI-generated. By combining spatial and temporal analysis of voice signals, the system effectively detects synthetic speech with high accuracy.

Experimental results demonstrate the robustness of the proposed approach in identifying deepfake voices using benchmark datasets. The integration of a real-time alert mechanism enhances proactive fraud prevention in banking communications.

The modular and scalable architecture allows seamless integration with existing financial systems. Future work will focus on improving multilingual support and enhancing robustness against evolving voice synthesis techniques.



Sample output of the Deepfake Detection

7.REFERENCES

[1] Y. Wu, P. Li, and L. Li, "Detecting synthetic speech using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[2] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 1008–1014.

[3] G. Lavrentyeva, S. Novoselov, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in Proc. Interspeech, 2017, pp. 82–86.

[4] J. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329–5333.

[5] X. Zhang, Z. Zhao, and Y. Qian, "Hybrid CNN-RNN models for real-time audio classification," IEEE Access, vol. 11, pp. 45678–45689, 2023.

[6] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, and M. Todisco, "ASVspoof 2021: Benchmarking spoofed and deepfake speech detection," in Proc. IEEE ICASSP, 2021, pp. 6369–6373.

[7] A. Oord et al., "WaveNet: A generative model for raw audio," in Proc. IEEE Speech Synthesis Workshop, 2016.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[9] T. Kinnunen, M. Sahidullah, H. Delgado, et al., "The ASVspoof 2017 challenge: Assessing spoofing countermeasures for automatic speaker verification," Computer Speech & Language, vol. 47, pp. 1–20, 2018.

[10] S. Singh and R. Patel, "AI-based fraud detection mechanisms in banking systems," Journal of Cybersecurity and Digital Forensics, vol. 5, no. 2, pp. 45–56, 2022.

[11] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, pp. 130–153, 2015.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

