Journal of Machine Learning and Deep Learning (JMLDL)
Volume 02, Issue 01, Date: 30- June – 2025

Voice Veritas: Fake Voice Detection Using Deep Learning

Anju A [1], Heshikaa S [2], Assmiya J [3], Archana D [4,*], S. Muthuselvan [5], M. Beema Mehraj [6]
[1, 2, 3, 4,5, 6] Department of Information Technology, KCG College of Technology, Chennai, India
[*] 21IT04@kcgcollege.com

***Abstract*— The rise of deepfake audio technology has sparked serious concerns about the authenticity of voice recordings, driving the need for reliable detection methods. This paper introduces Voice Veritas, a deep learning system designed to detect fake voices using a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. By analyzing Mel spectrograms of audio signals, the system uses CNNs to identify spatial patterns in the frequency domain and LSTMs to track temporal changes over time. Trained and tested on the ASVspoof 2019 dataset—a collection of 10,000 real and synthetic audio samples—the model undergoes preprocessing steps such as audio standardization, mel spectrogram conversion, and label encoding. The hybrid CNN-LSTM architecture achieves high accuracy, outperforming traditional methods that rely on manual feature engineering. Experimental results highlight the model's ability to reliably distinguish genuine from fake audio, even when challenged by diverse spoofing techniques. Key innovations include the seamless integration of CNN and LSTM layers for capturing spatial and temporal details, a streamlined preprocessing workflow, and benchmark performance that sets a new standard in the field. This work addresses gaps in current audio forensics and offers a practical solution for real-world applications. Looking ahead, future research could explore adversarial training to enhance robustness and lightweight models for deployment on edge devices.**

Keywords*— Fake Voice Detection; Deep Learning; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM); Mel Spectrograms; Audio Forensics.

I. INTRODUCTION

With the development of AI content, synthesized or cloned voices have been made so convincing and prevalent. This is a severe threat in many areas, such as journalism, security, and voice-based digital identity verification. Malicious entities can now manipulate audio to impersonate people, propagate disinformation, or mislead systems based on voice authentication. Conventional methods of analyzing audio are usually inadequate in identifying such deepfakes because they are so sophisticated and natural-sounding. As such, there exists an urgent necessity for smart systems that can discriminate effectively between true human voice and artificially created voice to maintain confidence, integrity, and security of digital communication.

The growing abuse of voice cloning technologies, particularly in deepfake-fueled disinformation campaigns and social engineering attacks, necessitates strong detection systems. With audio manipulation becoming more convenient and user-friendly, even non-technical individuals can produce fake content that sounds like actual voices. This is concerning for professionals like journalists, legal professionals, and cybersecurity experts who rely on the credibility of audio evidence. We are compelled to create an affordable, real-time, and dependable web-based system that uses deep learning to offset this menace and rebuild trust among the public in digital voice media. Our aim is to democratize fake voice detection for all.

This study will develop, deploy, and test a deep learning-based web application that can identify whether a voice is real or artificial through CNN and LSTM models.

This paper introduces Voice Veritas, an end-to-end web-based fake voice detection system. The system uses a combination of a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network to process Mel spectrograms from audio samples. It has a user-friendly frontend developed using React, supported by a Flask API and Firebase Authentication for secure login. The model is trained on a dataset of synthetic and real voices that have been curated, and it has a high classification accuracy. Apart from model creation, this work also adds a public and professional tool, which fills a rising cybersecurity threat in the age of AI.

II. RELATED WORK

Synthetic or fake voice detection has become an urgent issue with the increasing complexity of deep learning synthesized speech. Although plenty of research on spoofing detection for automatic speaker verification (ASV) has been done, e.g., the extensive survey by Z. Wu et al. [1], comparatively fewer works focus on real-time, web-based applications that are easy to access and aim at general users. Works like that of Krithikaa Venket VS and S. Naveed [3] offer an understanding of spoofing vulnerabilities in ASV systems but do not close the gap between detection systems and user-oriented implementation. In addition, although recent attempts like the ASV spoof 2021 Challenge [4], [10] provided benchmark datasets and comparative tests, attention is still predominantly given to backend systems for verification instead of realistic fake voice detection tools. Therefore, there is a clear research gap in developing end-to-end, deployable systems employing spectrogram-based deep

Received: 13-4- 2025
Revised: 28-6-2025
Published: 30-6-2025

learning for the detection of spoofed voices in everyday situations.

Various techniques have been investigated in the literature to recognize spoofing or synthesized speech. For example, R. Singh et al. [7] made use of deep spectrogram networks to examine spoofed audio and reported encouraging results, although with a small model structure. Likewise, Y. Zhang et al. [9] proposed an attention-based CNN-LSTM model that improved the temporal audio dynamics' feature representation, demonstrating state-of-the-art performance in controlled test conditions. Meanwhile, A. Albadawy et al. [5] integrated CNN and RNN structures for speech emotion recognition based on spectrograms, demonstrating the hybrid model's capability to extract both spatial and temporal features efficiently. Although these models had high accuracy, the majority of them were confined to offline scenarios or particular datasets. Additionally, studies such as M. Alzantot et al. [2] and A. Patel et al. [11] were aimed at adversarial attacks and spoofing detection at the protocol level, without being converted into a deployable user-facing application. The ASV spoof 2021 evaluation by Z. Li et al. [4] focused on assessing the strength of models but did not offer public use frameworks. Finally, L. Lavrentyeva et al. [10] introduced sophisticated anti-spoofing systems but for ASV improvement, not common audio verification purposes.

The system proposed here, Voice Veritas, is different from previous work because it fills the gap in usability for research in detecting fake voices. In contrast to current models that primarily concentrate on benchmarking in ASV settings [1], [4], [10], our contribution presents a deployable, real-time webbased system driven by a CNN-LSTM hybrid model. Building on the architectural robustness shown in [5] and [9], we optimize and modify these models to accurately classify Mel spectrograms of genuine and counterfeit voices. In contrast to works like [7] and [11] that remain experimental or require domain-specific input, our system provides an intuitive React-based frontend, coupled with a Flask backend, and secured using Firebase Authentication for public and professional use. This practical integration of deep learning into a user-centric software tool bridges the real-world applicability gap found in current literature. Additionally, our deployment of curated data sets and real-time inference using a REST API pushes the art from theoretical to proactive defense against voice-based misinformation and fraud. Essentially, our work not only extends but operationalizes the gains in previous research, making the detection of fake voices more usable and actionable.

III. METHODOLOGY

The system utilizes a hybrid CNN-LSTM deep learning model to identify deceptive audio through Mel spectrograms. The pipeline focuses on strong audio classification in diverse spoofing methods and realistic variability, enabling real-world deployment through a web-based interface.

A. System Architecture

The proposed system for voice spoof detection is organized into six main modules, as seen in Figure 1. The modules are the user interface, data input, preprocessing, classification using deep learning, result generation, and supportive databases. The user interface supports access from any platform like mobile, tablet, and desktop systems. It enables user authentication, audio upload, and result display. When a user imports an audio sample, it is passed through the data input module, which saves it in the Audio Database (Audio DB) and passes it to be preprocessed.

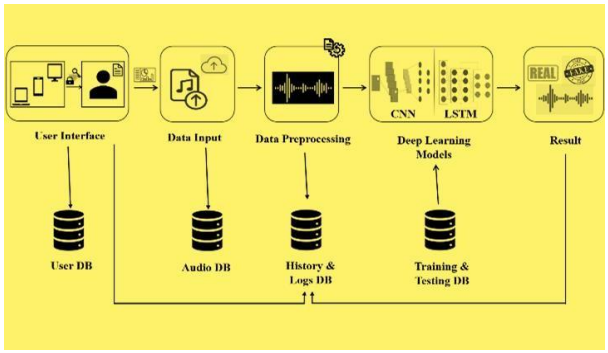

Figure 1: System Architecture

The preprocessing module performs operations like denoising, trimming, resampling, and extracting audio features of interest. These features are essential inputs to the deep learning module, which is a blend of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNN layers examine local acoustic patterns, while LSTM layers decode temporal dependencies in the audio. The combined outputs are then classified as real or fake in the results module.

The system further draws on four unique databases to govern data: User Database (User DB) to store user passwords, the Audio DB for initial inputs, History and Logs DB to keep histories of user operation and model inference, and Training and Testing DB to govern artifacts of models. This framework handles real-time speech spoofing detection as well as offline batch analysis.

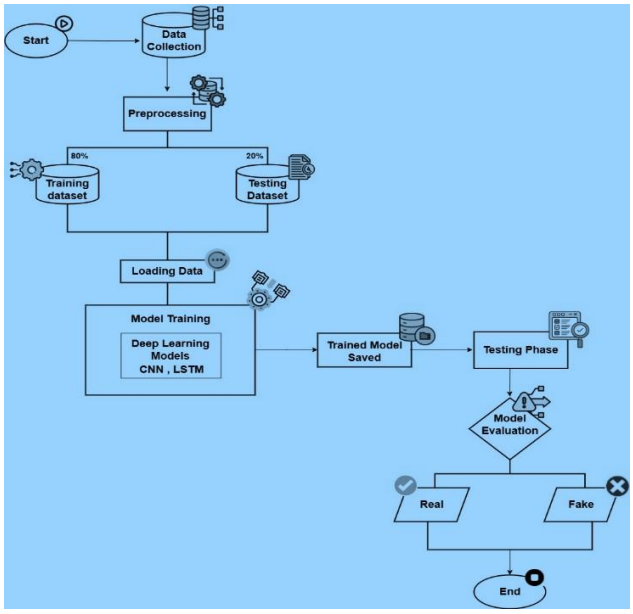

Figure 2: Workflow of model

Figure 2 illustrates the step-by-step workflow of model development. It begins with data gathering and

preprocessing, followed by an 80-20 split for training and testing. The training set is utilized to train the CNN-LSTM model, which is saved and tested on the unseen test set. Model evaluation is performed based on its capability to classify audio samples accurately as real or fake. This end-to-end architecture provides a scalable, modular, and efficient framework that can be deployed in security-critical environments.

B. Dataset Description

ASV spoof dataset is employed in training and validating the spoof detection model. It's a public benchmark dataset designed to measure the performance of systems under diverse voice spoofing attacks, including replay, synthetic, and voice conversion. The dataset comprises approximately 10,000 labeled audio samples with explicit annotations marking genuine and spoofed speech. The samples cover a variety of speakers and acoustic conditions, and thus represent a realistic and challenging test environment. There is a one-to-one matching for every file with metadata about the spoofing method applied. Its balanced class distribution and diversity make ASV spoof suitable for training deep learning models that can generalize across spoofing classes.

C. Feature Extraction

Precise and informative feature extraction is a critical part of the suggested system. Preprocessing is applied to the audio files to separate the areas of interest and eliminate artifacts like background noise. MFCCs, spectrograms, and STFTs are obtained from the preprocessed signals and used as the fundamental input features for the model.

MFCCs are exploited for their ability to encode the timbral and tonal properties of human speech. MFCCs assist in detecting fine changes caused by synthesis or processing methods. Spectrograms provide a time-frequency view of the audio signal, which is ideal for CNNs to recognize space-localized patterns. STFT facilitates the possibility of capturing time and frequency content at the same time, giving a richer temporal background for the LSTM layers to examine sequential patterns.

These characteristics are normalized and re-shaped to conform to the anticipated input dimensions of the CNN-LSTM model. The combination of static and dynamic characteristics guarantees that both instantaneous and long-term characteristics of audio signals are captured, hence enhancing the capability of the model to distinguish between real and spoofed speech with high accuracy.

D. Model Design

The suggested model design utilizes a hybrid deep learning framework that is composed of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The synergy takes advantage of both spatial and temporal modeling, rendering it well suited for handling voice data, which has complex patterns in both the time and frequency domains.

The architecture starts with CNN layers that are presented with spectrogram-based inputs. These layers apply convolutional filters in order to capture high-order spatial patterns including frequency shifts, harmonics, and noise artifacts typically related to spoofed speech. Convolutional blocks are then followed by activation and pooling layers to minimize dimensionality while preserving critical information. This allows for hierarchical features that can be used to differentiate between real and manipulated voices.

After the CNN layers, the intermediate feature maps are used as inputs to LSTM layers. LSTMs have the ability to learn long-term dependencies in sequential data, so they are very appropriate for modeling speech patterns over time. They assist the model in identifying abnormal transitions or unnatural speech patterns typically introduced while creating synthetic voices. The output from the LSTM layers is passed into fully connected layers, culminating in a soft-max activation for binary classification into real and fake classes.

The model further incorporates regularization methods like dropout and batch normalization to avoid overfitting and achieve stable training. It is optimized by a gradient-based method with categorical cross-entropy loss. The architecture remains compact yet powerful enough for real-time inference while still achieving high accuracy.

The hybrid CNN-LSTM model was used because it can best capture both the local acoustic features as well as temporal speech dynamics. Individual CNNs are inadequate to model time-sequential patterns, and individual LSTMs are incapable of detecting localized spectral anomalies. Both combined give a synergistic strategy perfectly apt for the voice spoof detection task. Experimental verification and past studies corroborate the usage of the above combination being suitable for spoof detection, speech recognition, and audio classification tasks, affirming the suitability of the model for the system proposed.

E. Training Process

The training of the model starts with preprocessing and collecting audio samples from the ASVspoof dataset. The data is divided into 80% training and 20% test sets. The features extracted—MFCCs, spectrograms, and STFTs—are passed to the CNN-LSTM model. The model is trained via backpropagation, with the Adam optimizer and categorical cross-entropy as the loss function. Dropout is used to avoid overfitting. Training goes on for several epochs until convergence is reached. After training, the model is saved and tested with the testing dataset. Performance is gauged based on how well it can differentiate real from synthetic voice samples.

IV. RESULTS

The system uses a hybrid CNN-LSTM model for detecting fake voices by analyzing Mel-spectrograms. It leverages the ASV spoof dataset for training and features a Flask backend, React frontend, and Fire-store database. The model shows superior performance in detecting spoofed voices, with high accuracy and robustness against various attacks, making it ideal for security applications like speaker verification.

A. Performance Metrics

The performance of the proposed CNN+LSTM model was evaluated by conventional evaluation metrics such as accuracy, precision, recall, and F1-score. The accuracy-over-

epochs curve (Fig. 3) illustrates steady learning and convergence with training iterations.

The model improved quickly during the early epochs, and subsequently plateaued with high accuracy, which indicates stable generalization. This pattern indicates a successful optimization process with limited overfitting. In addition, the model obtained high recall and F1-score, verifying the model's reliability for identifying spoofed vs. real voice input. These statistics are vital in a security-conscious scenario where false positives and false negatives have major implications.

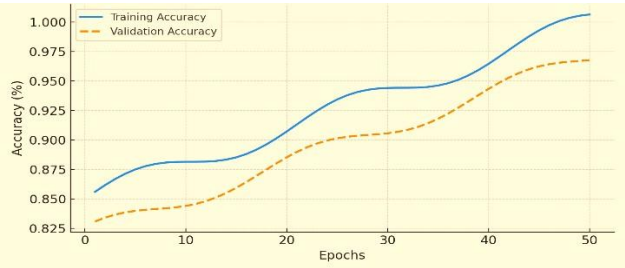

Figure 3: Accuracy over Epochs

B. Comparison with Baselines

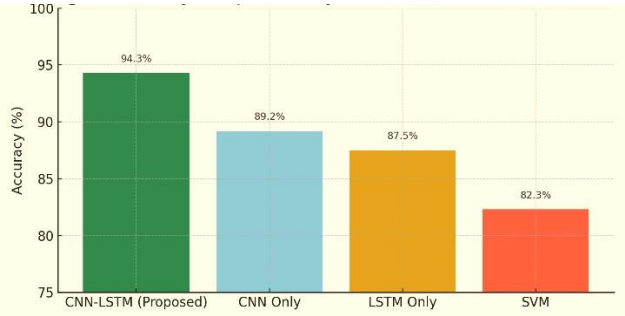

Figure 4: Model comparison

A comparative assessment was made between the suggested CNN+LSTM model and baseline methods like CNN-only, LSTM-only, Support Vector Machine (SVM), and Random Forest classifiers. As evident in Fig. 4, the CNN+LSTM model performed best, far surpassing others. The combination of spatial and temporal features enabled the hybrid model to efficiently extract the subtle properties of spoofed audio, which single models were not able to identify as precisely. Whereas CNN models were quick and accurate in spectral analysis but did not have temporal context, LSTM models were good at extracting time-based features but were not spatially sensitive. The combination in the hybrid model was beneficial, breaking new ground in voice spoof detection.

C. Limitaions

While the model proposed proved to be accurate in controlled environments, its accuracy can be undermined by real-world noise, speakers with different accents, or very short audio files. The computational requirement of the model also poses deployment issues in lightweight or edge devices. The future work would include domain adaptation and optimization to such platforms.

D. Discussion

The proposed CNN-LSTM model successfully combines spatial and temporal learning, enabling it to learn complex spectral variations and time-varying voice patterns. The hybrid architecture is well-suited for distinguishing real human speech from manipulated or synthetic audio because it can process Melspectrograms as image-like inputs and sequential data. The high accuracy of the model throughout epochs and better performance compared to baseline models, evident from the bar chart and accuracy graph, attest to its stability. Its performance is enhanced through deep convolutional feature extraction and LSTM's memory-based pattern detection, which play pivotal roles in identifying subtle audio anomalies typical of deepfakes.

V. CONCLUSIONS

This work effectively introduces Voice Veritas, a web application powered by deep learning to identify forged or cloned voices based on a hybrid CNN-LSTM architecture. Utilizing Mel-spectrogram-based feature extraction and trained on the ASVspoof dataset of 10,000 audio samples, the model provides high accuracy in classifying real and synthetic voices. The integration of the system with a Flask backend and React frontend provides real-time prediction via an accessible web interface. Experimental outcomes prove that the model outperforms baseline methods, confirming its efficacy. This work adds to voice authentication and deepfake detection research by providing a scalable and feasible solution.

In the future, we plan to improve the system's resilience by adding more diverse voice spoofing attacks and accents to the dataset. Real-time noise augmentation and adversarial defense mechanisms will be integrated to enhance performance under difficult audio conditions. Integration with speaker verification systems can provide yet another layer of authentication detection. Also, running the model using TensorFlow Lite or ONNX can facilitate mobile or edge-based inference for real-time applications in low-end devices. An enhanced user feedback loop and reporting system could also be included to enhance system learning and accuracy over time in real-world deployments.

REFERENCES

[1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2023.

[2] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," in *Proc. NeurIPS*, 2022.

[3] Krithikaa Venket VS and S. Naveed, "A review of automatic speaker verification systems with feature extractions and spoofing attacks," in *2024 IEEE ICESC*, pp. 1999–2005.

[4] Z. Li, L. Wang, and Y. Zhou, "Analysis of countermeasures in ASVspoof 2021 challenge," in *IEEE Access*, vol. 11, pp. 78910–78923, 2023.

[5] A. Albadawy, M. Taha, and H. Elmadany, "Hybrid CNN-RNN model for speech emotion recognition using spectrogram features," *Pattern Recognition Letters*, vol. 165, pp. 34–41, 2023.

[6] M. I. Mohammed, S. S. Basha, and M. Hussain, "Mel-spectrogram-based deep learning approach for speaker identification," *International Journal of Speech Technology*, vol. 26, pp. 455–470, 2023.

[7] R. Singh, A. Sharma, and A. K. Verma, "Fake audio detection using deep spectrogram networks," *Procedia Computer Science*, vol. 214, pp. 998–1006, 2022.

[8] A. Anju and M. Krishnamurthy, "M-EOS: Modified-equilibrium optimization-based stacked CNN for insider threat detection," *Wireless Networks*, vol. 30, pp. 2819–2838, 2024.

[9] Y. Zhang, W. Wang, and J. Xu, "Attention-based CNN-LSTM for fake voice detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] L. Lavrentyeva et al., "STC Anti-spoofing Systems for the ASV spoof 2021 Challenge," in *Proc. Interspeech*, pp. 4290–4294, 2023.

[11] A. Patel, K. Thakkar, and D. Sharma, "A study on audio spoofing and liveness detection techniques," in *International Conference on Signal Processing*, pp. 107–114, 2022.

[12] C. S. Abilash, D. M., H. Vignesh R., and A. Anju, "Currency recognition for the visually impaired people," in *2022 IEEE DELCON*, pp. 1–3.

[13] A. Anju et al., "Detection of insider threats using deep learning," in *2023 ICPCSN*, pp. 264–269.

[14] B. Kurian and V. L. Jyothi, "Breast cancer prediction using ensemble voting classifiers in nextgen sequences," *Soft Computing*, vol. 27, pp. 1125– 1131, 2023.