**COMPARATIVE ANALYSIS OF DEEPFAKE DETECTION MODELS ON DIVERSE VOICES**

Siddharth Bhorge, Ruta Sapate, Samruddhi Raut,

Depart of E&TC Engineering , Depart of E&TC Engineering , Depart of E&TC Engineering ,

Vishwakarma Institute of Technology, Vishwakarma Institute of Technology, Vishwakarma Institute of Technology,

Pune,India Pune,India Pune,India

[siddharth.bhorge@vit.edu](mailto:sakshi.rathod22@vit.edu) [ruta.sapate22@vit.edu](mailto:ruta.sapate22@vit.edu) [samruddhi.raut22@vit.edu](mailto:samruddhi.raut22@vit.edu)

Sakshi Rathod,

Depart of E&TC Engineering ,

Vishwakarma Institute of Technology,

Pune,India

[sakshi.rathod22@vit.edu](mailto:sakshi.rathod22@vit.edu)

*Abstract -*Cloned voices have become more common as artificial intelligence in speech synthesis has advanced, making it harder to identify the difference between authentic and fraudulent audio. A hybrid deep learning model that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks is used in this paper to offer a Cloned Voice Detection System. In order to capture both synthetic artifacts and human speech characteristics, the suggested method collects important audio components such as spectrograms, Fast Fourier Transform (FFT), and Mel-Frequency Cepstral Coefficients (MFCCs). For effective model training, the retrieved features go through preprocessing, normalization, and dataset partitioning. While the BiLSTM layers gradually pick up sequential dependencies, the CNN layers examine spatial patterns in the audio characteristics Evaluations using confusion matrix analysis and AUC-ROC curves show that the method is effective in differentiating between real and cloned voices, and the trained model is used for real-time inference, allowing automated detection of deepfake audio.

*Keywords — CNN-BiLSTM, MFCC, FFT, Spectrogram, Deep Learning, Speech Forensics, Voice Authentication, Synthetic Speech Recognition.*

1. INTRODUCTION

The rapid advancements in artificial intelligence (AI) and deep learning have significantly improved the quality of speech synthesis and voice cloning technologies. Text-to-speech (TTS) systems powered by artificial intelligence (AI) are capable of producing incredibly lifelike human voices, which makes them valuable for uses like accessibility tools, audiobook narration, and virtual assistants. These developments have brought up security issues, though, since bad actors can use them to produce deepfakes that sound like real people, which can result in identity theft, false information, and voice spoofing attacks. Conventional approaches to voice authentication and speaker verification depend on statistical models and manually developed feature extraction methods, which frequently miss minute artifacts in cloned voices. In order to distinguish between authentic and fake audio, contemporary deepfake detection methods use deep learning techniques to analyze both spectral and temporal features. This research offers a Cloned Voice Detection System employing a hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) model to improve detection accuracy. In order to capture both the natural vocal characteristics and the artificial distortions found in cloned voices, the suggested approach collects important elements from audio signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), Fast Fourier Transform (FFT), and Spectrograms. The model is successful at identifying minute distinctions between actual and artificial intelligence-generated sounds because CNN layers examine spatial patterns in these features and BiLSTM levels gradually learn sequential dependencies.

An actual and cloned voice dataset is preprocessed and separated into training, validation, and testing subsets in order to train and assess the model. The preprocessing stage involves normalization, noise reduction, and feature extraction to ensure that the input data is optimized for learning. The extracted features—Mel-Frequency Cepstral Coefficients (MFCCs), Fast Fourier Transform (FFT), and Spectrograms—are stored in NumPy (.npy) files for efficient access and reduced computational overhead during training. The model training process utilizes Binary Cross-Entropy Loss as the loss function, which is well-suited for binary classification tasks such as distinguishing between real and fake voices. The Adam optimizer is applied to boost training efficiency and convergence speed by dynamically altering learning rates. Learning rate scheduling and early stopping are used to minimize overfitting and enhance model performance. The training is conducted in mini-batches (batch size = 32 or 64) over multiple epochs (typically 10-20 epochs), ensuring that the model learns both low-level and high-level patterns in the audio features. Post-training, the final trained model is saved as deepfake\_voice\_model.h5 and deployed for real-time inference. During deployment, the system accepts a new audio file as input, processes it through the same feature extraction pipeline, and passes the extracted features to the CNN-BiLSTM model for classification. The model then outputs a probability score, determining whether the audio sample is real or fake. A threshold-based decision is applied to label the input accordingly. To ensure scalability and real-world applicability, the deployed model can be integrated into various voice authentication systems, forensic analysis tools, and cybersecurity frameworks. Additionally, it can be continuously updated with new datasets to improve its ability to detect evolving deepfake techniques.

To further improve detection accuracy and resilience against adversarial attacks, future developments might incorporate ensemble learning techniques, dataset augmentation, and adaptive learning. By offering a scalable and effective framework for identifying AI-generated speech and reducing the risks associated with deepfake audio manipulation, this research adds to the ongoing efforts in speech forensics, cybersecurity, and AI-based fraud detection.

1. **LITERATURE REVIEW**

A hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for identifying phony speech recordings was developed and evaluated. The goal of the study was to address the growing concern about AI-generated speech that can mimic human voices, which presents security and privacy issues. Preprocessing methods including segmentation and normalization were employed on a balanced dataset that included 5,889 actual and 5,889 false speech samples. Extensive training, validation, and hyperparameter adjustment were performed on the CNN-LSTM model. With a 99.2% accuracy, 99.2% F1 score, 99.4% recall, and 99.0% precision, the experimental findings showed great efficacy. These results demonstrate how hybrid deep learning models may be used to reduce the possibility of digital communications fraud involving fake voices[1]. Recent developments in deep learning have greatly enhanced the creation of synthetic voice and speech recognition. The difficulties with speaker diarization and synthetic speech are the main emphasis of this study, specifically with regard to differentiating between authentic and fraudulent voices in group discussions. Several datasets, including Urban-Sound8K, Conversational, AMI-Corpus, and FakeOrReal, were integrated into a deep learning-based system that was suggested. Four key components comprised the approach: (1) a CNN-based fake audio detection model with 94% accuracy; (2) speaker diarization using Natural Language Processing for text conversion (93% accuracy) and a Recurrent Neural Network (RNN) for speaker labeling (80% accuracy, 0.52 DER); and (3) a speech-denoising module using Multilayer Perceptron and Convolutional Neural Networks (CNN) with 93% and 94% accuracy, respectively.These findings demonstrate the effectiveness of deep learning in combating synthetic voice-based threats and enhancing speech analysis applications[2]. As part of the EUCINF (EUropean Cyber and INFormation) project, this study presents a deep learning-based system for detecting deepfake audio. Three transformation methods were used to turn the raw input audio into spectrograms: Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT), along with auditory-based filters like Mel, Gammatone, linear filters (LF), and Discrete Cosine Transform (DCT). Three deep learning approaches were investigated: (1) baseline CNN, RNN, and C-RNN models; (2) transfer learning using computer vision models like ResNet-18, MobileNet-V3, EfficientNet, and DenseNet-121; and (3) audio pre-trained models like Whisper, Speechbrain, and Pyannote, with extracted embeddings classified using a Multilayer Perceptron (MLP). The best-performing models were fused to improve performance, achieving an Equal Error Rate (EER) of 0.03 [3]. As artificial intelligence (AI) develops quickly, deepfake media production has improved thanks to Generative Adversarial Networks (GANs), which presents significant threats in industries like politics, journalism, and law. Conventional deepfake detection techniques use recurrent networks for temporal analysis and convolutional networks for geographical analysis. This paper suggests SFormer, a unique transformer-based model for both spatial and temporal deepfake detection, in order to overcome their shortcomings. SFormer makes use of the Swin Transformer to improve generality across various manipulation techniques and lower computer complexity. Tests on several deepfake datasets, including FF++, DFD, Celeb-DF, DFDC, and Deeper-Forensics, showed excellent performance, with accuracy rates of 100%, 97.81%, 99.1%, 93.67%, and 100%, respectively. The outcomes demonstrate how reliable and successful SFormer is in comparison to other deepfake detecting techniques[4]. Widely utilized in consumer IoT applications, voice-driven devices (VDDs) like Google Home and Amazon Alexa are susceptible to logical access (LA) attacks such voice conversion and text-to-speech (TTS) synthesis. These flaws can be used by attackers to get around security and access private systems, such as home automation and banking, without authorization. A unique Extended Local Ternary Pattern (ELTP) feature descriptor that captures algorithmic artifacts and dynamic vocal tract features in synthetic and converted speech is proposed in this study to address this problem. To improve the ability to distinguish between actual and spoof voices, the ELTP features are merged with Linear Frequency Cepstral Coefficients (LFCC). To distinguish between real and fake signals, a Deep Bidirectional Long Short-Term Memory (DBiLSTM) network is trained using these combined properties[5]. By allowing for the very accurate identification and classification of voice, music, and ambient noises, deep learning has greatly improved the classification of audio signals. Before deep learning models process audio signals, they are first modified using methods such as spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and wavelet decomposition. Convolutional neural networks (CNNs) are used for speaker identification and speech recognition, recurrent neural networks (RNNs) are used to record temporal audio patterns, autoencoders are used for feature extraction, transformers are used to analyze temporal and frequency-based features, and hybrid models, which combine deep learning and conventional classifiers, are used. Together, these strategies improve audio classification systems' precision and effectiveness[6]. Advances in deep learning have made deepfake generation much better, which has raised privacy and security issues. Detecting AI-generated bogus audio, such as replay assaults, voice conversion, and text-to-speech (TTS), is the goal of audio spoofing detection. The lack of generality in traditional machine learning techniques results in inaccurate detection. In order to improve feature extraction from mel-spectrograms, we suggest a deep-layered model that combines VGGish with a Convolutional Block Attention Module (CBAM). This architecture captures spatial and channel-based correlations to properly classify audio as real or fake. Our model's efficiency in identifying audio deepfakes was demonstrated by its Equal Error Rate (EER), which was 0.52% for Physical Access (PA) attacks and 0.07% for Logical Access (LA) assaults when tested on the ASVspoof 2019 dataset[7]. This paper provides a comprehensive survey on the latest advancements in deepfake generation and detection methods. Deepfake content, including highly realistic images, audio, and videos created using artificial intelligence, poses serious threats to national security, democracy, and personal privacy. The paper categorizes deepfake generation methods into face swapping and facial reenactment, while detection approaches primarily rely on feature-based and machine learning techniques. However, deepfake detection still faces challenges such as the continuous improvement of deepfake generation techniques, the scarcity of high-quality datasets, and the absence of standardized benchmarks. Future research directions focus on developing efficient, robust, and systematic detection methods along with improved datasets to enhance the reliability of deepfake detection systems[8].This study examines the methods used to create and identify audio deepfakes, going over broad deepfake principles, important approaches for creating audio deepfakes, and detection tactics. Support Vector Machines (SVMs), Decision Trees (DTs), Convolutional Neural Networks (CNNs), Siamese CNNs, Deep Neural Networks (DNNs), and hybrid CNN-RNN architectures are examples of detection techniques, along with statistical analysis and media consistency checks. According to performance evaluation across research, DT had the lowest accuracy at 73.33%, while SVM had the greatest at 99%. Siamese CNNs demonstrated a 55% improvement in the tandem detection cost function (t-DCF), while Deep-Sonar's Equal Error Rate (EER) ranged from 2% to 12.24%[9]. With uses ranging from helping with speech impairments to facilitating financial fraud and fake news, fake voices have emerged as a significant problem in social media, cybersecurity, and forensics. The creation of extremely organized and realistic speech that accurately replicates human voices is made possible by developments in AI-driven technologies like as Google Audio LM. A deep learning and machine learning-based approach to phony speech detection is suggested as a countermeasure. To identify whether a voice is synthetic or real, mel-frequency cepstral coefficients (MFCCs) are taken out as features and categorized using several machine and deep learning models. The efficiency of this method in differentiating between real and phony voices is confirmed by experimental results[10]. Deepfake videos, which frequently include both audio and visual alterations, raise serious questions regarding the validity of internet media. Conventional detection techniques concentrate on intra-modal artifacts, whereas multi-modal analysis is necessary for real-world deepfakes. In order to detect deepfakes, this research suggests AVoiD-DF, an Audio-Visual Joint Learning technique that takes use of audio-visual discrepancies. The framework consists of a Cross-Modal Classifier for identifying discrepancies, a Multi-Modal Joint-Decoder for fusion, and a Temporal-Spatial Encoder for feature extraction. A new benchmark dataset, DefakeAVMiT, spanning a variety of forgery strategies across modalities, is presented to improve evaluation. DefakeAVMiT, FakeAVCeleb, and DFDC experimental results show that AVoiD-DF outperforms state-of-the-art techniques and exhibits strong generalization across various forgery strategies[11].

It is becoming difficult for both humans and automatic speaker verification (ASV) systems to discern between real and phony voices due to the advent of voice phishing, or vishing, made possible by deep learning-based speech synthesis. This study presents BTS-E, a framework that assesses the relationship between breathing, talking (speech), and silence sounds in an audio clip in order to improve deepfake detection. This attribute is used for deepfake detection because text-to-speech (TTS) systems have a hard time mimicking genuine breathing. The efficacy of this method is confirmed by extensive tests on the ASVspoof 2019 and 2021 datasets, which show a 46% improvement in classifier performance[12]. ASVspoof Challenges have prompted study into detecting techniques due to the increasing dangers posed by deepfake audio. However, the Speech Deepfake (DF) subset, which includes a variety of spoof audio sources, continues to be a challenge for state-of-the-art models. In order to improve generality and resilience, this study suggests a unique detection architecture. To enhance representation, the method combines hand-crafted features with learnt embeddings. Furthermore, it uses insights from several deepfake generating methods to formulate the training procedure as a bi-level optimization issue. Experimental findings show that, when used as a stand-alone system without the need for ensemble modeling or data augmentation, this approach provides the best detection performance documented in the literature[13].The challenge of synthetic voice and spliced audio spoofing is addressed by improving existing countermeasures. Traditional methods treat detection as a binary classification problem (bonafide vs. spoof), but this work extends Res2Net with a Conformer block to better capture local acoustic patterns. The proposed SE-Res2Net-Conformer architecture significantly improves detection performance for logical access spoofing, according to experimental results on the ASVspoof 2019 dataset. The paper also suggests a novel approach to audio splicing detection, which shifts the focus from identifying entire spliced segments to detecting their precise boundaries. Unlike previous signal processing-based methods, this approach uses deep learning to achieve more accurate and effective splicing detection[14]. The ethical, legal, and societal ramifications of artificial media material are examined in this essay, with a focus on journalism, education, and security. Given the proliferation of software tools that may create and alter digital voices, the ability to identify phony voices is essential for stopping fraudulent activity. The paper suggests identifying artificially manufactured sounds by utilizing convolutional neural networks (CNNs) with Mel spectrograms. To find the best CNN architecture for language-independent detection, supervised experiments were carried out utilizing voice samples from several datasets. The greatest accuracy ratings were 98% for WaveFake, 94% for ASV, and 99% for the FoR dataset. The model's accuracy was 98% for FoR, 92% for ASV, and 96% for WaveFake when it was trained on all datasets at once and tested on separate datasets[15].

1. **METHODOLOGY**

**REFERENCE**

1. Oyucu, S., Çelımlİ, D. B. Ü., & Aksöz, A. (2024, November). Fake Voice Detection: A Hybrid CNN-LSTM Based Deep Learning Approach. In 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-6). IEEE.
2. Wijethunga, R. L. M. A. P. C., Matheesha, D. M. K., Al Noman, A., De Silva, K. H. V. T. A., Tissera, M., & Rupasinghe, L. (2020, December). Deepfake audio detection: a deep learning based solution for group conversations. In 2020 2nd International conference on advancements in computing (ICAC) (Vol. 1, pp. 192-197). IEEE.
3. Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024, September). Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. In 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2) (pp. 1-5). IEEE.
4. Kingra, S., Aggarwal, N., & Kaur, N. (2024). SFormer: An end-to-end spatio-temporal transformer architecture for deepfake detection. Forensic Science International: Digital Investigation, 51, 301817.
5. Arif, T., Javed, A., Alhameed, M., Jeribi, F., & Tahir, A. (2021). Voice spoofing countermeasure for logical access attacks detection. IEEE Access, 9, 162857-162868.
6. Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access, 11, 106620-106649.
7. Kanwal, T., Mahum, R., AlSalman, A. M., Sharaf, M., & Hassan, H. (2024). Fake speech detection using VGGish with attention block. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 35.
8. Zhang, T. (2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5), 6259-6276.
9. Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio deepfake approaches. IEEE Access, 11, 132652-132682.
10. Kilinc, H. H., & Kaledibi, F. (2023, October). Audio deepfake detection by using machine and deep learning. In 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 1-5). IEEE.
11. Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18, 2015-2029.
12. Doan, T. P., Nguyen-Vu, L., Jung, S., & Hong, K. (2023, June). BTS-E: Audio deepfake detection using breathing-talking-silence encoder. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
13. Li, M., Ahmadiadli, Y., & Zhang, X. P. (2023, September). Robust Deepfake Audio Detection via Bi-Level Optimization. In 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP) (pp. 1-6). IEEE.
14. Wang, L., Yeoh, B., & Ng, J. W. (2022, December). Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 115-119). IEEE.
15. Valente, L. P., de Souza, M. M., & Da Rocha, A. M. (2024, May). Speech Audio Deepfake Detection via Convolutional Neural Networks. In 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-6). IEEE.