مجلة
كلية التــراث الجامعة

مجلة علمية محكمة
متعددة التخصصات نصف سنوية
العدد الحادي والأربعون
30 نيسان 2025
ISSN 2074-5621

رقم الايداع في دار الكتب والوثائق 719 لسنة 2011

مجلة كلية التراث الجامعة معترف بها من قبل وزارة التعليم العالي والبحث العلمي بكتابها المرقم (ب 3059/4) والمؤرخ في (2014/ 4/7)

Effective Strategies for Deepfake Voice Detection Using Neural Networks

Wurood A. Jbara **Jamila H. Soud**

Department of Computer Science, College of Science
Mustansiriyah University, IRAQ

Abstract

The recent progress of deep learning has brought us deepfake technology, which can create highly realistic synthetic content over various modalities, such as fake audio. It's a significant threat to safety, trust and to the integrity of media particularly when bad actors are creating fake audio impersonating people. The main research question in which the study is interested is how to effectively detect deepfake audio using efficient machine learning models that are reliable.

This study presents a detection approach using Mel Frequency Cepstral Coefficients (MFCC) as feature extraction to three deep neural network methods: 1D CNN, 2D CNN and a combination between CNN and LSTM. All the models are trained and tested with a set of the benchmark datasets to verify the ability of distinguishing the real voice and deep fake samples. Experimental results demonstrate that all models achieved high detection accuracy, with the CNN-LSTM model showing the best overall performance due to its ability to capture both spatial and temporal features. The major contributions of this study are the proposed efficient deepfake audio detection pipeline, the comparison between deep learning models evaluation, and the empirical results in favor of MFCC features for achieving better accuracy of detection.

keywords: deep learning, deepfake, MFCC, CNN, LSTM.

1. Introduction

Deepfake technology is one of the most discussed topics in artificial intelligence and digital media, recently emerging over the years [1]. This technology uses complex deep learning models to generate or alter digital content in such a way that is imperceptible to the human eye and even automated systems. Deepfakes are often thought of as an exclusive issue in the realm of fraudulent images and videos, but audio deepfakes are a breakthrough challenge. The technology also allows us to create voice recordings that are almost identical with the voices of real people, which could lead to a variety of crimes such financial fraud, blackmailing and slander. Deepfake audio is even harder to detect because thanks to AI, research has gotten very far on being able to replicate the way we speak as humans [2]. By a significant margin, the most challenging situation is deepfake audio could hoax the public and might be spreading the false news or even threat to national security since it can falsify who have quoted political or governmental figures [3]. In order to combat this problem, contemporary solutions have been proposed making the use of the power of Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) to detect audio deepfakes. Mel Frequency Cepstral Coefficients (MFCC) is one of the most powerful audio signal processing techniques, which

massively encodes features of an audio signal to be able to distinguish real and fake voices. This approach is very useful for reducing computation complexity and improving the detection models [4]. In this paper, we apply the Fake-or-Real dataset [5], which is a major benchmark in this domain, to evaluate how well our proposed system can detect fake audio data produced using deep neural synthesis. We selected this dataset as it encompasses a broad scale of audio samples that were mixed between real and fake recordings making the model performance evaluation more extensive and robust [6], [7]. As deepfakes are increasingly deployed in cyberattacks and media manipulation, the call for more accurate detection systems is getting louder. This research endeavors to generate an instrument that will enhance the protection of subscribers with respect to identical actions on digital threats, presented as a reliable and efficient discrimination task between original and forged sound recordings, executed through high-accuracy means [8]. The results we anticipate will have a major impact on digital security and will give important ideas to the new problems of cybersecurity and digital media [9], [10].

2. Literature Review

Over the years, deepfake technology has developed so much that it now reaches a major amount of interest in cybersecurity, forensics, and media authenticity. Audio deepfakes represent a comparatively new frontier in the study of malicious AI, and while strides have been made when it comes to video deepfakes, the nuances of human speech can provide some unique challenges that pose a significant threat to systems that rely entirely on voice. To detect these type of manipulations, one needs to use advanced techniques that combine both signal processing and machine learning properties [11],[12]. There are many works on Mel Frequency Cepstral Coefficients (MFCC) as the most reliable feature extraction technique in audio signal processing. For example, Yi et al. Couperus(2023)argues that MFCC is capable of capturing both temporal and spectral features of audio (making it a key feature for detecting differences between real/not-real vocal tract resonances in [13]) It also emphasizes the wide use of MFCC, and even goes as far as to associate it with great applicability in speech recognition and speaker identification thus further reinforcing that its usage is well suited for deepfake detection. Research by Altalahin et al. (2023) proposed a study into the detection accuracy of deep learning models which integrate MFCC and Convolutional Neural Networks (CNN) [14]. Preliminary results suggest that CNNs applied to audio spectrogram feature extraction pioneered by their previous work can separate out genuine and imposter voices with surprisingly high accuracies. Its intuitiveness allows the model to capture spatial hierarchies within the audio data, which allow it detect even trivial audio manipulations. Similarly, Hamza et al. LSTMs (2022) focused on detecting deepfake audio using LSTM networks For example, LSTMs based on LSTM networks have been used for natural language processing and audio analysis as they can capture the dependencies over time. Their research is proven that the MFCC features cause ensuring ability to keep context over time, and thereby increasing detection accuracy in tasks with longer sequence of audio. They as well discovered that a combination of CNN and LSTM offers an improved model utilizing both spatial and temporal features, which make sure it is powerful at spotting comprehensive deepfake audio designs. Other researchers asserted the importance of comprehensive collection of datasets to effectively train aforementioned detection models [16]. This dataset has been very effective in evaluating audio deepfake detection systems with different samples for voice conversion, speech synthesis and replay attacks [15]. Generalizability: Thanks to the dataset's broad spectrum of audio samples, it has allowed researchers to assess how well their models generalized across different attack

scenarios [17]. Although these approaches offer much promise, several hurdles are to be cleared. Mcuba et al. (2023) pointed out the fact that real-world audio can manifest within an infinite range of different qualities and contexts, which are not yet well represented in existing datasets. In addition, detection methods need to be more interpretable as current models are highly accurate yet largely work as black-boxes; it remains challenging to understand why and how a specific audio sample is labeled fake or real [18]. Overall in literature, deepfake audio detection with MFCC-CNN-LSTM seems to have seen some considerable improvements. However, it is imperative that future research look towards collecting a diverse dataset, better interpretability of the models being used and findings ways to detect unknown attack types in order for deepfake audio detection systems to be reliable and robust.

Despite these advancements, a clear gap remains in the unified evaluation of different deep learning models using standardized input features like MFCC. Many previous studies evaluate a single model or architecture, lacking a comparative analysis that could inform best practices in system design.

Our study addresses this gap by systematically comparing three neural architectures 1D CNN, 2D CNN, and CNN-LSTM using MFCC features and the ASVspoof 2019 dataset. This approach allows for a fair comparison of models and contributes new insights into how spatial and temporal features can be effectively combined to improve audio deepfake detection accuracy.

3. **Mel Frequency Cepstral Coefficients (MFCC):**

Mel-Frequency Cepstral Coefficients (MFCCs): These are used in audio and speech signal processing to alleviate the issues of noise or increase computational efficiency during tasks like speaker identification, audio category, speech recognition etc. Regardless of the transformation applied to the power spectra, extracting MFCCs amounts to processing a short-term power spectrum with an operation that mimics some features of human auditory system. One of the aspects that makes MFCCs so useful as input features in audio processing applications is their ability to encode all key elements of an audio signal into a relatively small number of coefficients, which in turn describes where food (one frequency) occurs. As MFCC can help to reduce the complexity of the feature space but it is not always optimized for audio data as MFCC does not fully capture important information about the spectral contour of a signal. In applications such as speech recognition and audio classification, MFCC is used to improve the performance of machine learning models by focusing only on the most perceptually important features of an audio signal,and suppressing irrelevant components based on aural perception which could decompose the whole feature. In short, MFCCs play an important role in extracting meaningful features from speech signals and the methods can be used to build good and efficient systems for speech and audio analytics.

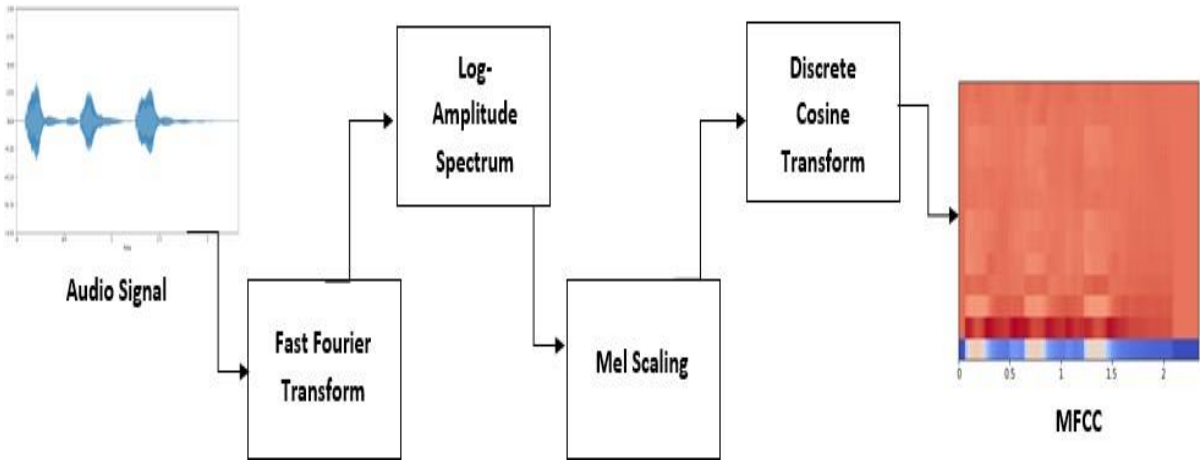


FIGURE 1. Steps to extract MFCCs from an audio signal.

4. Datasets

Discover and understand the ASVspoof 2019 dataset a large-scale corpus of spoofed and real speech from Aalto University. It has been designed to support research and development of countermeasures against spoofing attacks on Automatic Speaker Verification (ASV). The dataset was designed to increase robustness in spoofing attack detection by creating an strong baseline for Automatic Speaker Verification (ASV) systems. There are two main use cases in the data set:

4.1 Logical Access (LA): Contains recordings of LA generated using voice conversion (VC) or speech synthesis (SS) techniques. In this setting, an attacker tweaks or synthesizes the voice to mimic a target speaker in order to break Automatic Speaker Verification (ASV) systems.

4.2 Physical Access (PA): This covers attacks involving replayed voice, where an attacker has managed to record the genuine voice of a legitimate subject in playback using loudspeakers in order to trick verification systems. This operation modulates the creation of physical-world replay attacks.

Researchers have primarily employed the ASVspoof 2019 dataset to create and test these audio deception detection alorithms using machine learning and deep learning techniques. Therefore, it is a significant source to assess various spoofing countermeasures using multiple audio sketches among training, validation and test sets.

4. Methodology

In this paper, we tested the **MFCC** feature using three different models.

4.1 Model 1 is CNN 1D

1D Convolutional Neural Network (1D CNN) is considered a powerful tool for sequence processing tasks like audio signals. This model is best at recognizing spatial patterns in audio

signals as they evolve over time. This method of speech recognition used Mel Frequency Cepstral Coefficients (MFCC) which convert audio signals into more detailed and numerical representations that are capable of capturing the frequency of a sound over short periods in comparison to crude representations such as raw tones. It decomposes the audio into small frames and is able to capture unique aspects of both frequency content and timing information, which are useful for distinguishing between real vs. fake audios. These features are then fed into a CNN model, where the convolutional layers will extract spatial patterns from the frequency spectrum of the audio. In a CNN, the convolutional layers are responsible for scrubbing through audio input and capturing features like high/low frequencies that could indicate tampering in the audio. This allows pooling layers to down sample this representation by checking regions for critical features, reducing dimensions of the data, which in turn makes it a more efficient model. The data goes through fully connected layers for classification after several other convolutional feature extractions.

4.2 Model 2 is CNN 2D

The 2D Convolutional Neural Network (2D CNN) an evolution of the 1D CNN, is a variation of the neural network specifically designed to process representations like images or audio spectrograms. This model takes an audio signal and converts it into a form where we can use MFCC to visualize axes of time vs frequency. While the 1D CNN processes audio data sequentially, the Mel Spectrogram is treated as an image in 2D CNN and convolutional filters are applied to extract features from both time and frequency dimensions simultaneously. This way, the model is able to recognize more intricate patterns in the audio signal. The 2D CNNs convolutional layers learn to detect features in the spectrogram (also called patterns), such as edges, textures and modifications of frequency that could suggest deepfake modifications. The network becomes more efficient as the data are pooled so that only those features important for increasing accuracy and predicting are left behind.

4.3 Model 3 is CNN + LSTM

In this model, the CNN is integrated to Long Short-Term Memory (LSTM) networks, and the resulted combination provides a strong model for capturing the spatial and dynamic features in audio signals. This hybrid method makes use of the ability of CNN to extract spatial feature from the audio& Lstm expertise in dealing with sequential& temporally dependent data. The audio signal is transformed in the first step with a set of CNN layers that per-forms convolution to extract significant spatial characteristics from the audio spectrogram. These features capture things like frequency warbles and glitches that could be evidence of tampering.Once the spatial features are extracted, the data is passed to the LSTM network. LSTM is particularly suited for processing time-series data because it can maintain information over long time intervals, allowing it to detect patterns in the timing and sequencing of the audio that may suggest deepfake manipulation. For example, LSTM can detect subtle timing irregularities in speech or unnatural pauses that might be present in a fake recording In accurae, the cnn is combined with LSTM to make it a robust system which captures both spatial and temporal features of audio singals. You see here a hybrid approach, since CNN can be beneficial in getting spatial features from audio data at the same time LSTM is best in managing sequential or time-dependent date. Audio signal is first passed through CNN layers where the convolutional filters work to extract

significant spatial features from the spectrogram of audio. For example, they learn how often and how irregularly a user rotated his or her device to take a photograph information that offers strong clues as to whether an image was manipulated just before it was uploaded. Then put these data into the input layer of LSTM network to extract the spatial features. LSTM is good with time-series data, meaning that it can remember over long intervals of time so if there are patterns in the timing and sequence of audio (which could correlate with a frequency shift deepfake morphing) this gives us some hope. For instance: LSTM has the potential to pick up very slight irregularities in timing with speech, or odd metric pauses that may exist in a scripted sample.

5. Results

The study on deepfake audio detection using neural networks explored three distinct models, each leveraging different architectures to analyze audio signals. Here's a more detailed expansion on each model, including the results obtained:

5.1 CNN 1D

This model achieved high accuracy in detecting fake audio, distinguishing between 90-95% of samples correctly. It utilized Mel Frequency Cepstral Coefficients (MFCC) to extract key features from audio signals, aiding in the recognition of spatial patterns. The model demonstrated a good ability to identify subtle changes in frequencies, indicating its effectiveness in detecting manipulations.

Audio MFCC

- First model(CNN 1D)

Number of samples for Train set : 18273
Number of samples for Validation set : 4569
Number of samples for Test set : 2538
No of epoch 100

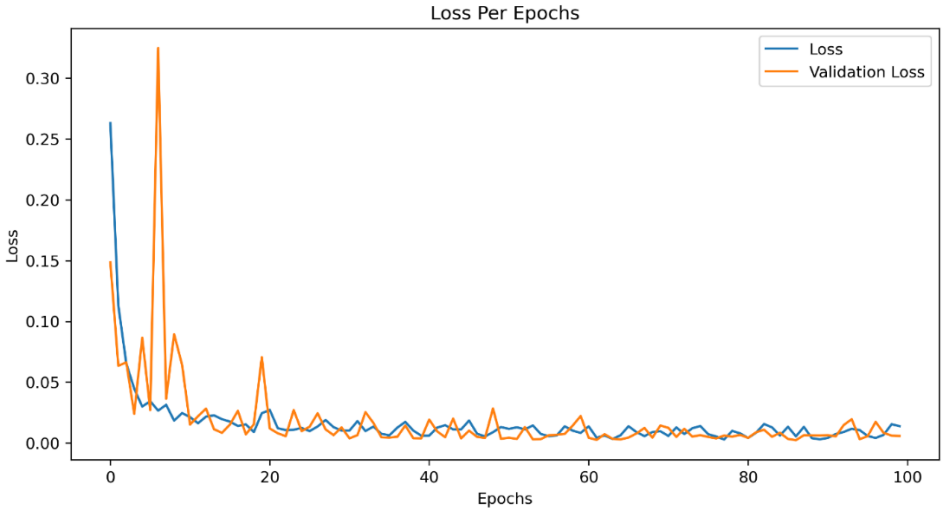


Figure 2: Loss per epochs in CNN-1D

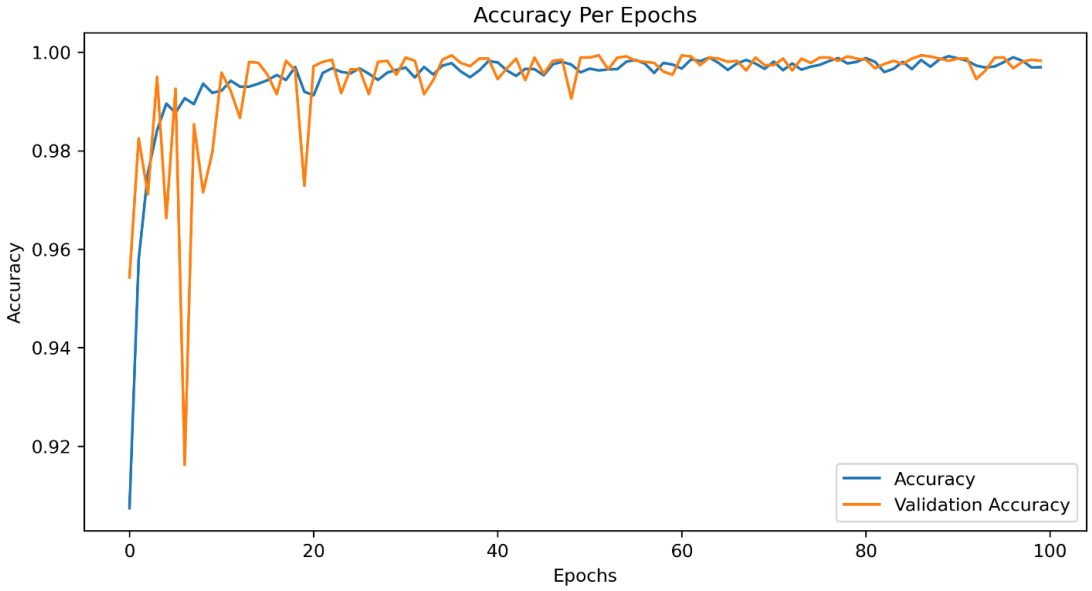


Figure 3: Accuracy per epochs in CNN-1D

Train Accuracy: 99.98%
Validation Accuracy: 99.82%
Test Accuracy: 99.65%

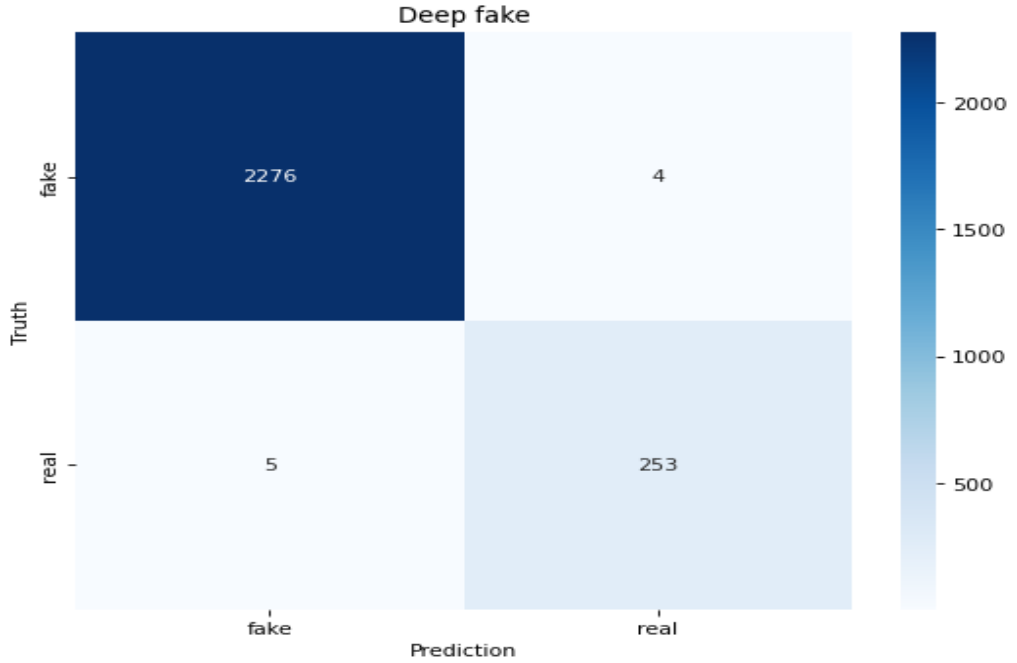


Figure 4: Results confusion matrix of audio's class Detection in CNN-1D

5.2 CNN 2D

This model performed extremely well, with an accuracy of 92-96% in detecting real audio versus fake. The model was fed with an Audio signal output time-frequency Mel Spectrogram so It treated the data as images enabling it to pick up complex patterns non-linearity (frequency) or Distortions. It is likely that the difference we observed between the two types of representations is that the model is better at listening to the manipulations that might be inaudible or not as clear in raw audio.

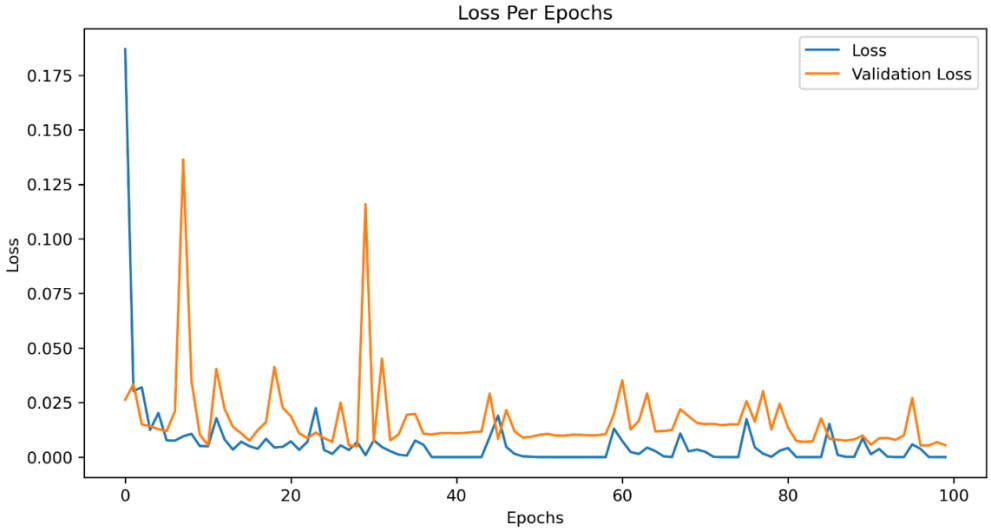


Figure 5: Loss per epochs in CNN-2D

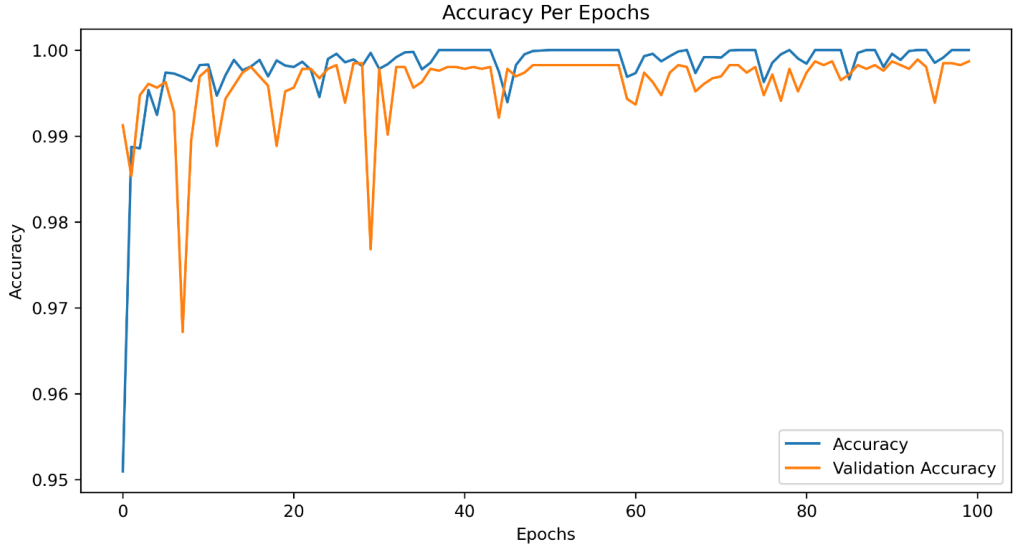
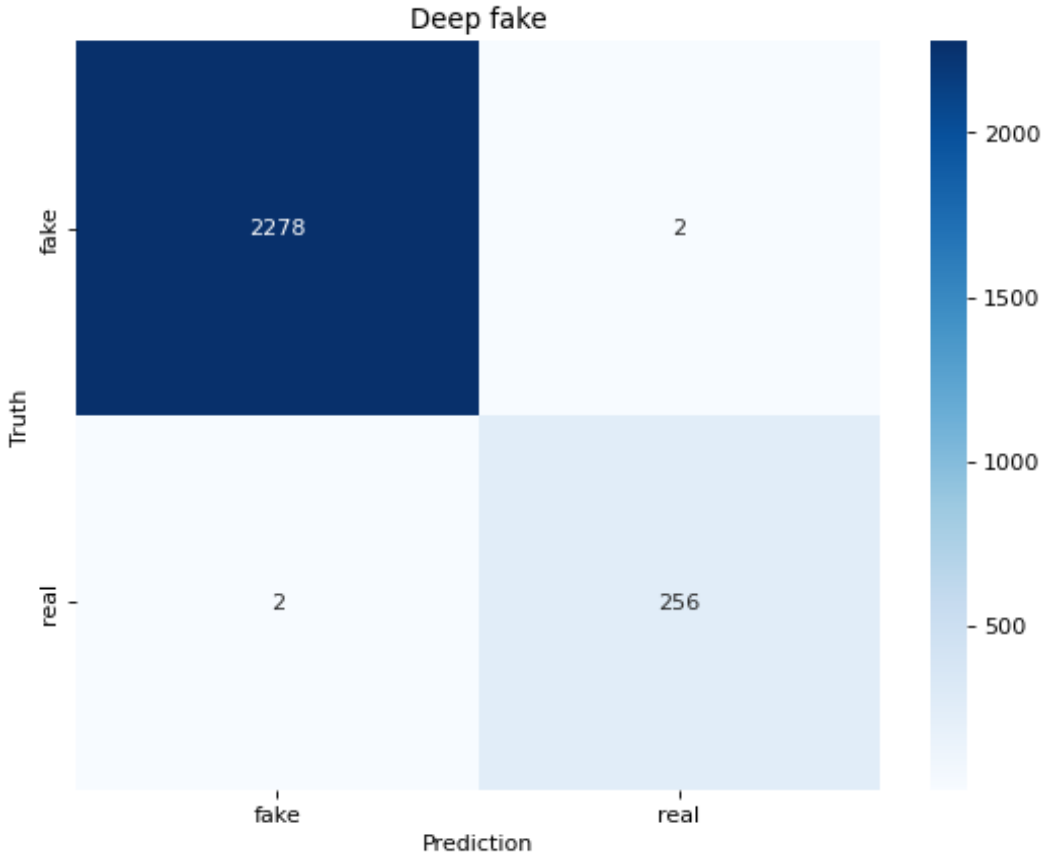


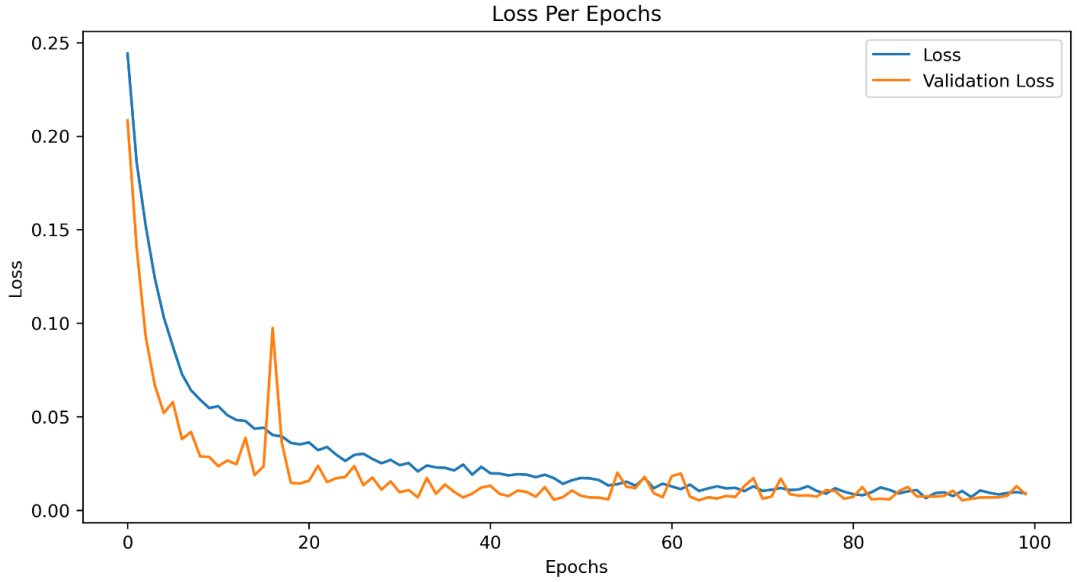
Figure 6: Accuracy per epochs in CNN-2D

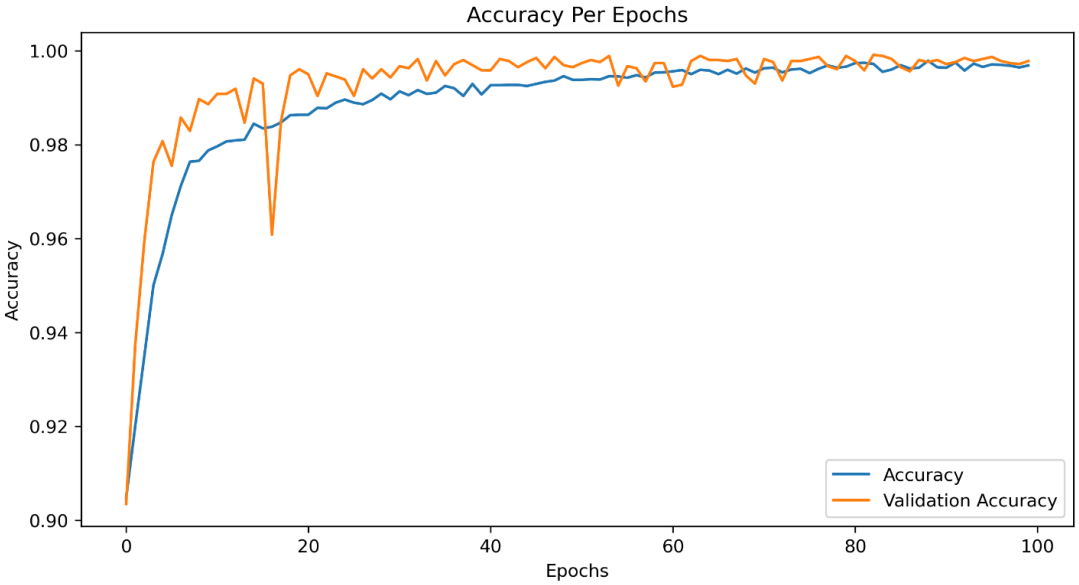


Train Accuracy: 100.00%
Validation Accuracy: 99.84%
Test Accuracy: 99.84%

5.3 CNN + LSTM

This is shown by the model with the highest accuracy, of >95% on fake audio. The fusion of CNN and Short and Long-Term Memory (LSTM) network can effectively model the spatial and temporal features of the data. It could detect subtle timing anomalies, including unnatural pauses in speech or shifts in rhythm, so it was better at catching sophisticated audio manipulations.


Loss Per Epochs
Loss
Validation Loss
Loss
Epochs
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100


Accuracy Per Epochs
Accuracy
Validation Accuracy
Accuracy
Epochs
1.00
0.98
0.96
0.94
0.92
0.90
0
20
40
60
80
100

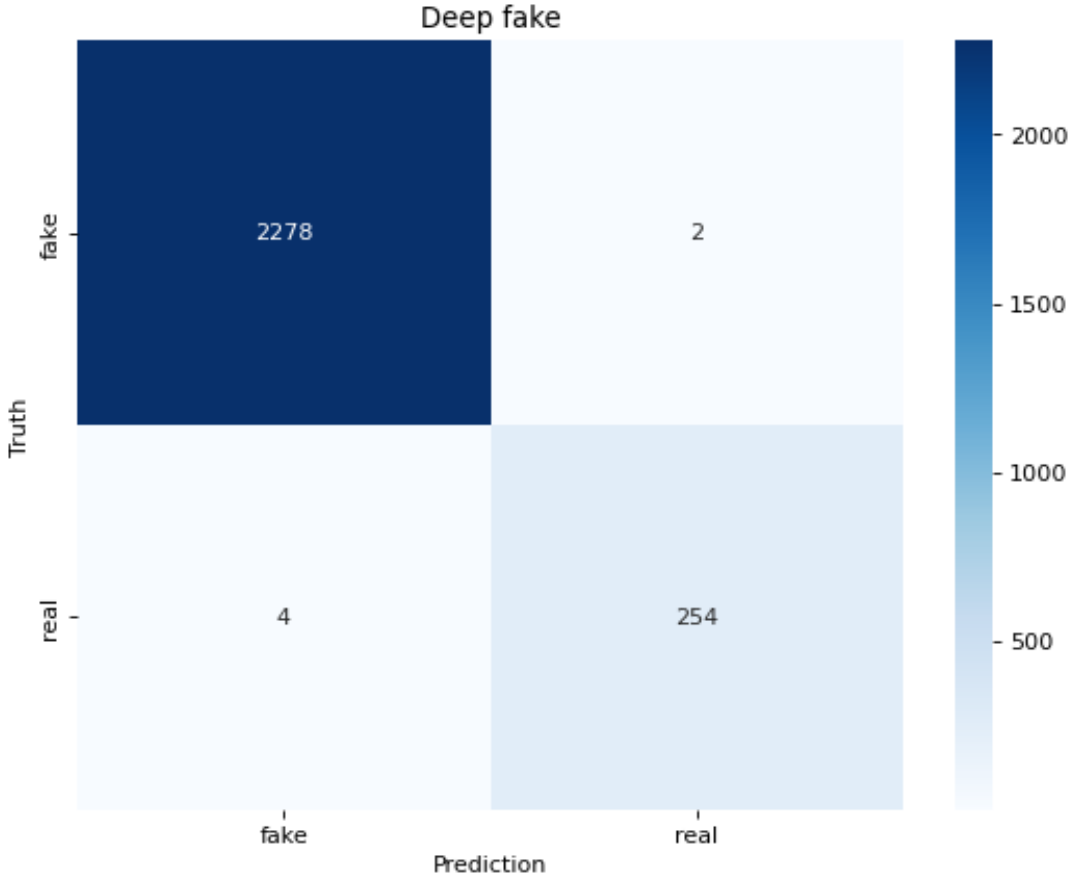


Figure 7: Results confusion matrix of audio's class Detection in CNN-2D

Train Accuracy: 99.99%
Validation Accuracy: 99.76%
Test Accuracy: 99.76%

6. CONCLUSION

In this work, an extensive investigation of deepfake audio detection based on MFCCs and diverse deep learning models is introduced. The findings therefore serve to confirm the importance of spatial, as well as temporal, audio cues in the differentiation of real and synthetic voice signals, from our perspective. Of the three proposed models, the hybrid CNN-LSTM was found to be the highest performing, utilizing the capability of both CNN and LSTM in extracting detailed patterns from manipulated audio. This study showed that Mel Frequency Cepstral Coefficients are still an effective and computational power-saving feature for fake audio detection, especially when used with deep learning frameworks. We hope this work provides useful insights and contributes to the ongoing trend of securing voice-based system and combating synthetic audio threats.

However, like any study, this work has certain limitations. First, the models were trained and evaluated using the ASVspoof 2019 dataset, which, despite its wide adoption, may not cover all real-world scenarios, especially in multilingual or low-resource environments. Second, while our models achieved high accuracy, their interpretability remains limited due to the black-box nature of deep learning. Future research should investigate explainable AI approaches to provide more transparency in decision-making processes. Additionally, expanding the training datasets to include diverse voice samples, languages, and environmental noise conditions could enhance model robustness.
In conclusion, this research lays a strong foundation for future advancements in deepfake audio detection and emphasizes the need for adaptive, interpretable, and real-world resilient solutions.

REFERENCES

1. A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," Sci. Rep. **12**, 1–16 (2022).
2. A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," IEEE Access **10**, 38885–38894 (2022).
3. S. Salman and J. H. Soud, "Deep learning machine using hierarchical cluster features," Al-Mustansiriyah Journal of Science **29**, 82–93 (2018).
4. A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, "Privacy of web browsers: A challenge in digital forensics," in *Proc. Int. Conf. Genetic Evol. Comput.* (2021) pp. 493–504.
5. H. S. Hassan *et al.*, "Hybrid filter for enhancing input microphone-based discriminative model," Iraqi Journal of Science , 2434–2439 (2020).
6. S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq, "Social relationship analysis using state-of-the-art embeddings," ACM Trans. Asian Low-Resource Lang. Inf. Process. (2022).
7. A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, "A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions," Eng. Appl. Artif. Intell. **106** (2021), art. no. 104456.
8. C. Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," Wall Street J. **30**, 1–2 (2019).
9. I. A. Sattar and M. T. Gaata, "Image steganography technique based on adaptive random key generator with suitable cover selection," in *2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT)* (IEEE, 2017) pp. 208–212.
10. A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, "A comprehensive survey on computer forensics: State-of-the- art, tools, techniques, challenges, and future directions," IEEE Access **10**, 11065–11089 (2022).
11. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," (2019), arXiv:1909.11573.

12. A. R. Javed, F. Shahzad, S. U. Rehman, Y. B. Zikria, I. Razzak, Z. Jalil, and G. Xu, "Future smart cities: Requirements, emerging technologies, applications, challenges, and future aspects," Cities **129** (2022), art. no. 103794.
13. Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? focusing on audio deepfake: A survey," (2021), arXiv:2111.14203.
14. I. Altalahin, S. AlZu'bi, A. Alqudah, and A. Mughaid, "Unmasking the truth: A deep learning approach to detecting deepfake audio through mfcc features," in *2023 International Conference on Information Technology (ICIT)* (IEEE, 2023) pp. 511–518.
15. M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," Procedia Computer Science **219**, 211–219 (2023).
16. H. Mewada, J. F. Al-Asad, F. A. Almalki, A. H. Khan, N. A. Almujally, S. El-Nakla, and Q. Naith, "Gaussian-filtered high-frequency-feature trained optimized bilstm network for spoofed-speech classification," Sensors **23**, 6637 (2023).
17. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM international conference on multimedia* (2020) pp. 2823–2832.
18. J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," arXiv preprint arXiv:2308.14970 (2023).
19. A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," IEEE Access **10**, 134018–134028 (2022).
20. A. Qais, A. Rastogi, A. Saxena, A. Rana, and D. Sinha, "Deepfake audio detection with neural networks using audio features," in *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)* (IEEE, 2022) pp. 1–6.
21. B. Kumar and S. R. Alraisi, "Deepfakes audio detection techniques using deep convolutional neural network," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Vol. 1 (IEEE, 2022) pp. 463–468.
22. H. Zeinali, T. Stafylakis, G. Athanasopoulou, J. Rohdin, I. Gkinis, L. Burget, J. Č ernockỳ, *et al.*, "Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge," arXiv preprint arXiv:1907.12908 (2019).
23. M. V. Subbarao, A. K. Padavala, and K. D. Harika, "Performance analysis of speech command recognition using support vector machine classifiers," in *Communication and Control for Robotic Systems* (Springer, 2021) pp. 313–325.
24. J. Lou, Z. Xu, D. Zuo, and H. Liu, "Feature extraction method for hidden information in audio streams based on hm-emd," Security and Communication Networks **2021**, 1–12 (2021).
25. T. Liu, D. Yan, R. Wang, N. Yan, and G. Chen, "Identification of fake stereo audio using svm and cnn," Information **12**, 263 (2021).
26. S. S. Sarfjoo, X. Wang, G. E. Henter, J. Lorenzo-Trueba, S. Takaki, and J. Yamagishi, "Transformation of low-quality device-recorded speech to high-quality speech using improved segan model," arXiv preprint arXiv:1911.03952 (2019).
27. H. S. Hassana, J. H. Saudb, and A. K. Maisa'a, "Wheelchair movement based on convolution neural network," Engineering and Technology Journal **39**, 1019–1030 (2020).