Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation

Taiba Majid Wani
Sapienza University of Rome
Rome, Italy
majid@diag.uniroma1.it

Syed Asif Ahmad Qadri
National Tsing Hua University
Hsinchu, Taiwan
syedasif@m110.nthu.edu.tw

Danilo Comminiello
Sapienza University of Rome
Rome, Italy
danilo.comminiello@uniroma1.it

Irene Amerini
Sapienza University of Rome
Rome, Italy
amerini@diag.uniroma1.it

ABSTRACT

Audio deepfake detection is emerging as a crucial field in digital media, as distinguishing real audio from deepfakes becomes increasingly challenging due to the advancement of deepfake technologies. These methods threaten information authenticity and pose serious security risks. Addressing this challenge, we propose a novel architecture that combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) for effective deepfake audio detection. Our approach is distinguished by the feature concatenation of a comprehensive set of acoustic features: Mel Frequency Cepstral Coefficients (MFCC), Mel spectrograms, Constant Q Cepstral Coefficients (CQCC), and Constant-Q Transform (CQT) vectors. In the proposed architecture, features processed by a CNN are concatenated into two multi-dimensional features for comprehensive analysis, then analyzed by a BiLSTM network to capture temporal dynamics and contextual dependencies in audio data. This synergistic method ensures an understanding of both spatial and sequential audio characteristics. We validate our model on the ASVSpoof 2019 and FoR datasets, using accuracy and Equal Error Rate (EER) metrics for the evaluation.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Audio Deepfakes, Feature Concatenation, MFCC, CQCC, CQT, Mel spectrograms, CNN, BiLSTM

ACM Reference Format:
Taiba Majid Wani, Syed Asif Ahmad Qadri, Danilo Comminiello, and Irene Amerini. 2024. Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IHMMSec '24), June 24–26, 2024, Baiona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3658664.3659647

This work is licensed under a Creative Commons Attribution International 4.0 License.

IHMMSec '24, June 24–26, 2024, Baiona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0637-0/24/06
https://doi.org/10.1145/3658664.3659647

1 INTRODUCTION

As synthetic voice synthesis technology becomes more sophisticated, audio deepfakes have emerged as a significant tool for deception. This advancement makes distinguishing between genuine and fake audio increasingly complex [11]. Despite advancements in this field, the existing methodologies are encumbered by significant demands on computational resources [10]. Optimal methodologies should, therefore, focus on enhancing feature extraction techniques, as these are critical in the efficient and effective detection of audio deepfakes. Feature extraction plays a pivotal role in enhancing the precision and reliability of audio deepfake detection systems [9]. Techniques such Mel-frequency Cepstral Coefficients (MFCC) streamline audio into a dense form that highlights crucial speech features, enhancing its utility in automatic speech recognition. This concentrated representation is particularly adept at uncovering inconsistencies inherent to deepfake audio, as it focuses on phonetic details pivotal to genuine human speech [4]. The Constant-Q Cepstral Coefficients (CQCC) feature stands out in the domain of spoof speech detection, offering a robust representation of audio data that captures subtle cues often overlooked by other methods [5]. Additionally, spectrogram features, which provide a visual representation of the spectrum of frequencies in a sound signal as they vary with time, have been acclaimed for their ability to represent complex audio patterns, thereby enhancing the detection process, especially when combined with other features like fundamental ferquency (F0) information and Real Plus Imaginary Spectrogram features [21]. A notable approach [15], involved the extraction of 60-dimensional linear filter banks, which serve as critical parameters in identifying the authenticity of audio samples. These filter banks capture the essential traits of audio, facilitating a more subtle and detailed analysis. Furthermore, advancements in the field have led to the creation of specialised datasets, such as the H-Voice dataset [2] which was constructed by extracting entropy features from both real and fake audio. This innovative approach not only enriches the dataset but also enhances the performance of detection models by providing a rich ground for training and evaluation.

Recent developments in the area of deepfake technology necessitate a continual advancement in detection methodologies, as the techniques for creating deepfakes are rapidly evolving. The adoption of deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has revolutionised the process of feature extraction in audio analysis [20]

[6]. These advanced neural network architectures facilitate a more comprehensive and nuanced understanding of audio data, surpassing traditional methods in their ability to classify complex patterns. This combination of state-of-the-art deep learning methods with classical acoustic analysis constructs a formidable and sophisticated system for detecting deepfake audio [8]. Such a hybrid approach is instrumental in safeguarding the authenticity of digital media content, a critical need in an era where the veracity of digital information is increasingly challenged. This integrated framework not only enhances the accuracy of deepfake detection but also fortifies the reliability of digital media, maintaining its trustworthiness in a landscape where digital deception is a growing concern.

Now, in this study, we present a novel method for audio deepfake detection by utilising four distinct feature sets: Mel Frequency Cepstral Coefficients (MFCC), Mel spectrograms, Constant Q Cepstral Coefficients (CQCC), and Constant Q Transform (CQT) features extracted from two prominent deepfake datasets, Fake or Real (FoR) [14] and ASVspoof 2019 [16]. The two datasets ensure a diverse range of real and fake audio samples. Upon the extraction of these features, they are systematically fed into a Convolutional Neural Network (CNN) designed for the extraction of hierarchical features from the input data. After processing, the feature sets are concatenated into two multidimensional vectors, one with MFCC, CQT, and CQCC, and the other with MFCC, Mel spectrograms, CQT, and CQCC. The first set (MFCC, CQT, CQCC) offers a robust spectral and temporal analysis and the second set (MFCC, Mel spectrograms, CQT, CQCC) provides detailed frequency-time representations along with spectral and temporal analysis. These concatenated feature sets are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM, by virtue of its architecture, analyses the temporal dynamics and contextual dependencies inherent in the sequence of audio data. The methodology culminates in a series of experiments where each of the concatenated feature sets is tested using the BiLSTM network to assess the efficacy of this novel approach in identifying audio deepfakes.

The rest of the paper is organised as follows: Section 2 explores some of the prior research on audio deepfake detection. Section 3 details our methodology, while Section 4 presents the experimental results and analysis. The paper concludes with a summary of findings in Section 5.

2 RELATED WORKS

This section explores a wide range of research in the field of audio deepfake detection, a domain characterised by significant advancements in extracting audio features and leveraging complex models like CNN and LSTM. The literature reviewed highlights a variety of methodologies and the extensive analytical work conducted by researchers in creating effective detection strategies. This includes detailed analysis of audio features and the sophisticated use of neural networks.

A. Qais et. al., [13] focused on the feature extraction from synthetic speech signals using Mel-Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT), and Spectrogram parameterization to capture the distinctive characteristics. These features were used both individually and in concatenated form to train and test the CNN model under various configurations.

Hamza et. al., [4] utilised several machine learning algorithms and deep learning models like LSTM and VGG16 and various features, MFCC, spectral (roll-off point, centroid, contrast, bandwidth), raw signal (zero cross rate), and signal energy for the detection of audio deepfakes. VGG16 achieved 93% of accuracy on FoR-original and SVM achieved 98.83% of accuracy on the FoR-rerec dataset of FoR dataset.

Mittal et. al., [12] utilized both static and dynamic Constant Q Cepstral Coefficients (CQCC) for feature extraction, employing a range of classifiers including LSTM, LSTM with Time Distributed Wrappers, and a two-dimensional Convolutional Neural Network (2D CNN). The research demonstrated improved performance when static and dynamic CQCC features were combined, achieving an Equal Error Rate (EER) of 0.009 on the ASVspoof 2019 dataset.

Krishnan et al., [7] implemented a multi-path strategy, processing three sets of features, MFCC, LFCC, and Chroma-STFT, through distinct, dedicated convolutional neural network (CNN) paths. This approach was designed to learn temporal patterns within the respective features, effectively capturing both local and global structures. The outputs from these paths were then integrated, resulting in the achievement of an Equal Error Rate (EER) of 0.69% on the In-the-Wild dataset and 0.79% on the FoR dataset, respectively.

Chakravarty et. al., [3] employed a ResNet50 for feature extraction on audio Mel spectrograms, followed by Linear Discriminant Analysis (LDA) for dimensionality reduction and several machine learning algorithms like Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN) and Support Vector Machine (SVM), for classification of real and fake audios. The proposed method performed better than the traditional features extraction methods, Gammatone Cepstral Coefficients (GTCC) and Mel Frequency Cepstral Coefficients (MFCC) and achieved an accuracy of 99.7%.

Inspired by previous research, our research introduces a novel approach to audio deepfake detection, marked by three key contributions: the use of CNN for spatial feature extraction from audio signals, concatenation of diverse feature sets including MFCC, Mel spectrograms, CQT, and CQCC using normalization and principal component analysis (PCA), and the application of BiLSTM for final classification. This combination of CNN-based spatial feature extraction, advanced feature concatenation, and BiLSTM classification distinguishes our method from existing techniques, aiming to enhance detection accuracy and efficiency in identifying audio deepfakes.

3 PROPOSED METHODOLOGY

In this study, we propose a structured pipeline for detecting audio deepfakes shown in Figure 1. The proposed approach utilises different techniques for extracting features and employs two neural network models, CNN and BiLSTM. The methodology is systematically organised into five main steps explained in this section.

3.1 Preprocessing

Initially several steps were undertaken to prepare the audio data for subsequent feature extraction and analysis. Noise reduction

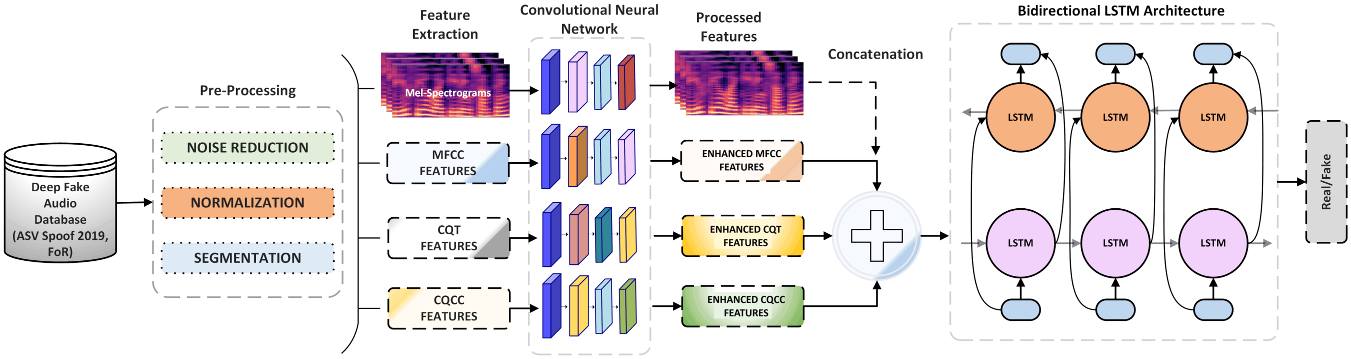

Figure 1: Proposed pipeline for the Audio Deepfake Detection

was implemented using a spectral gating method, which identifies and attenuates frequencies that fall below a certain threshold. This method is effective in reducing background noise and enhancing the clarity of the audio signals. Regarding normalization, the audio magnitudes were normalized to a consistent range between -1 and 1. This standardization ensures that the amplitude variations across different recordings do not bias the model's performance, allowing for a more accurate assessment of the audio content based on its features rather than its volume. To address the variability in audio sample lengths within our dataset, each audio file was segmented into uniform lengths of 3 seconds. This was achieved by truncating longer samples and padding shorter samples with zeros at the end. This approach ensures that each input to our model maintains a consistent temporal dimension, facilitating more reliable feature extraction and comparison across samples. The term 'segmentation' in our context refers specifically to this process of standardizing audio lengths, which is crucial for maintaining consistency in the inputs to our neural networks.

3.2 Feature Extraction

Four distinct feature sets are extracted from the preprocessed audio samples: Mel Feature Cepstral Coefficients (MFCC), Mel Spectrograms, Constant Q Transform (CQT) and Constant Q Cepstral Coefficients (CQCC).

3.2.1 Mel Feature Cepstral Coefficients (MFCC). Mel-Frequency Cepstral Coefficients (MFCC) are a prominent feature in audio analysis, particularly in speech processing and audio deepfake detection. Essentially, MFCCs are a representation of the short-term power spectrum of a sound, capturing its unique timbral qualities. Mathematically, the MFCC is derived by applying a Mel-scale filter bank to the log magnitude spectrum of the signal and then taking the discrete cosine transform of the log filterbank energies. The formula for calculating the i^{th} MFCC, C_i, is given as follows:

$$C_i = \sum_{j=1}^{N} \log(S_j) \cdot \cos\left(i \cdot (j - 0.5) \cdot \frac{\pi}{N}\right)$$

where S_j is the log energy in the j-th Mel-frequency bin, and N is the total number of Mel-frequency bins. In audio deepfake detection, the MFCC features are crucial as they encapsulate the dynamics of the vocal tract, which is significantly different in synthetic speech compared to natural human speech and is pivotal in distinguishing between real and fake audio. Furthermore, incorporating the first and second derivatives of MFCC (ΔMFCC and Δ^2MFCC) enhances the feature set, providing a more dynamic representation and add information about the rate of change in cepstral features.

3.2.2 Cepstral Q Transform (CQT). The Constant Q Transform (CQT) is a crucial tool in the field of audio analysis, unlike traditional Fourier transforms, it provides a time-frequency representation where the frequency bins are geometrically spaced, and the Q-factor (the ratio of the center frequency to the bandwidth) remains constant [1]. This characteristic makes the CQT exceptionally adept at analysing musical and complex audio signals, as it aligns more closely with human perception of pitch, especially useful for identifying the nuanced variations in pitch and timbre that are characteristic of deepfake audio. Mathematically, the CQT is defined by the equation

$$X_k(n) = \sum_{n=0}^{N-1} x(n) w_k(n) e^{-\frac{j2\pi kn}{Q}} \tag{1}$$

Here, $X_k(n)$ represents the CQT coefficient for the k-th frequency bin at time n. The function $x(n)$ denotes the audio signal, $w_k(n)$ is the window function applied to the k-th frequency bin, and Q is the quality factor, which determines the resolution and spacing of the frequency bins. The exponential term $e^{-\frac{j2\pi kn}{Q}}$ is a complex exponential that facilitates the transformation from the time domain to the frequency domain, similar to the role played by the Fourier transform but tailored for the logarithmic spacing of the CQT.

3.2.3 Constant-Q Cepstral Coefficients (CQCC). Constant-Q Cepstral Coefficients (CQCC) are a crucial component in the detection of audio deepfakes, building upon the strengths of the Constant-Q Transform (CQT) [12]. CQCC are obtained from CQT involving two key steps: logarithmic transformation and Discrete Cosine Transform (DCT). After computing the CQT as given in equation 1, the logarithm of the power spectrum is calculated. If $X_k(n)$ represents the output of the CQT for the k-th frequency bin at the n-th time frame, the logarithmic transformation is applied as follows:

$$L_k(n) = \log(|X_k(n)|^2) \tag{2}$$

where, $|X_k(n)|^2$ represents the power spectrum of the CQT, and $L_k(n)$ is the logarithmic power spectrum. The cepstral coefficients are then derived by applying the DCT to the logarithmic power spectrum. The DCT transforms the logarithmic power spectrum into the cepstral domain, capturing the envelope of the speaker's vocal tract. The formula for the DCT can be expressed as:

$$C_m = \sum_{k=1}^{K} L_k(n) \cdot \cos\left(\frac{\pi m}{K}(k - 0.5)\right) \tag{3}$$

In equation 3, C_m represents the m-th cepstral coefficient, and K is the total number of frequency bins in the CQT. The index k runs over all the frequency bins, and m typically ranges from 1 to a selected upper limit.

Combining these two steps, the CQCC feature set is formed, providing a robust representation of the audio signal's spectral characteristics[23].

3.2.4 Mel spectrograms. Mel spectrograms visually illustrate the spectrum of frequencies present in an audio recording and their variation throughout the duration of the file. Initially, the audio signal is transformed from the time domain to the time-frequency domain using the STFT represented in equation 4,

$$STFT(x(t)) = X(t, \omega) = \int x(t)w(t - \tau)e^{-j\omega t}dt \tag{4}$$

where $x(t)$ is the audio signal, $w(t - \tau)$ is a window function, and the exponential term $e^{-j\omega t}$ facilitates the conversion to the frequency domain. This step provides a detailed analysis of the frequency components within the signal across time. After obtaining the power spectrum from the STFT, it is mapped onto the Mel scale using the equation 5

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{5}$$

The Mel scale is a perceptual scale, designed to reflect the human ear's response to pitch, making the Mel Spectrogram especially adept at highlighting the peculiarities of human speech that are essential for identifying deepfake manipulations [19].

3.3 Feature Processing with CNN

In our proposed CNN architecture, we employ four convolutional layers to enhance feature extraction. The first layer uses 32 filters of size 3x3, followed by layers with 64, 128, and 256 filters of the same size, respectively. These layers progressively capture more complex patterns within the audio features like MFCC, Mel spectrograms, CQCC, and CQT. Each convolutional layer is accompanied by batch normalization to accelerate training and reduce overfitting, and ReLU activation functions to introduce non-linearities essential for deep learning. The dimensionality is reduced through max pooling operations, spaced after every two convolutional layers, using 2x2 pools to streamline the feature sets. This configuration not only ensures efficient learning but also helps in building a layered, hierarchical representation of the data crucial for detecting nuances in audio deepfakes. The final pooling layer's output is flattened and concatenated with other feature sets, preparing it for classification via the bidirectional LSTM.

3.4 Concatenation of CNN-Processed Features

Concatenation approach provides a comprehensive representation of audio signals and enhances detection accuracy. The process begins with normalizing the features to ensure equal contribution from each type and prevent any single feature from dominating the combined vector. Subsequently, two distinct multi-feature vectors are created: one combining MFCC, CQT, and CQCC, and the other additionally incorporating Mel spectrograms. This strategy not only captures a broad spectrum of audio characteristics, ranging from spectral details to cepstral features, but also facilitates a more thorough analysis due to the complementary nature of these diverse features. Principal Component Analysis (PCA) is then employed for feature selection. This approach effectively reduces the dimensionality of our feature sets while retaining the most significant features that are critical for deepfake detection. PCA is instrumental in enhancing model efficiency by focusing computational resources on the most informative aspects of the audio data. In our concatenation procedure, we ensured the temporal alignment of features, which was essential for accurately representing corresponding audio segments and maintaining the integrity of our analysis. We also emphasized feature selection to enhance our model's efficiency and reduce overfitting, selecting only the most pertinent features. Acknowledging the high-dimensional nature of this concatenated feature space, we adopted data augmentation techniques, crucial in boosting the model's ability to generalize. Finally, these temporally aligned features were processed using a BiLSTM model, adept at identifying temporal inconsistencies indicative of audio deepfakes.

3.5 Analysis with BiLSTM and classification

We utilize a three-layer Bidirectional Long Short-Term Memory (BiLSTM) network, where each layer consists of 128 hidden units. This structure allows the BiLSTM to effectively process audio data in both forward and reverse directions, enhancing its ability to capture crucial temporal nuances that are characteristic of deepfake audio. These nuances include inconsistencies like irregular speech rhythms and sudden shifts in tone that may not be detectable in the raw audio or through initial feature processing. By integrating concatenated feature vectors from various audio sources—MFCC, Mel spectrograms, CQCC, and CQT—the BiLSTM builds a robust representation of the audio landscape. This comprehensive view is pivotal for identifying subtle anomalies and temporal patterns within the audio data. The network's ability to remember and learn from long sequences significantly aids in distinguishing between authentic and fabricated content. Lastly, the classification layer of the BiLSTM utilizes these nuanced features to accurately classify audio samples as 'real' or 'fake', leveraging the depth of temporal insights gained through the bidirectional processing.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Datasets

4.1.1 ASVspoof 2019 dataset. The ASVspoof 2019 dataset [17] considers all three types of spoofing attacks—replay, speech synthesis, and voice conversion—in a single challenge. This database is divided into two scenarios: logical and physical access control, each

Table 1: Utterances in ASVspoof 2019 Dataset (LA)

Utterances	Training Set	Development Set	Evaluation Set
Bona fide	2,580	2,548	7,355
Spoofed	22,800	22,296	63,882

with its own distinct dataset. For our study, we have focused exclusively on the Logical Access (LA) scenario. In the LA condition, Text-to-Speech (TTS) and Voice Conversion (VC) algorithms are used to generate spoofed utterances. The database is organised into training, development, and evaluation sets, each encompassing multiple speakers. Furthermore, each set includes both bona fide (genuine) and spoofed utterances. The number of samples present in the LA scenario is detailed in Table 1.

4.1.2 Fake or Real (FoR) dataset. The Fake or Real (FoR) Dataset [14], essential for synthetic speech detection research, includes over 87,000 synthetic utterances and more than 111,000 real utterances, making it suitable for training complex deep learning algorithms. This dataset is available in four distinct versions: 1) The original version with unaltered speech source files, 2) A normalised version, balanced in gender and class and standardised in sample rate, volume, and channel number, 3) A 2-seconds version derived from the normalised version with truncated files, and 4) A rerecorded version designed to simulate scenarios such as phone calls or voice messages. Each of these versions comprises three sets: training, validation, and testing. For our study, we have merged all versions to utilise them as a single, comprehensive dataset, consisting of 53,000 real and 41,500 fake utterances, chosen randomly. We designed a custom split to ensure a balanced representation of real and fake audio samples. Specifically, the dataset was divided into 70% for training, 15% for validation, and 15% for testing.

4.2 Experimental Setup

In our deepfake detection study, both the CNN and BiLSTM models were configured with a batch size of 32, and training was conducted over 50 epochs. For optimization, the Adam optimizer was used, renowned for its efficiency in handling sparse gradients and adaptive learning rate capabilities. The cross-entropy loss function was used, a common choice in classification tasks due to its effectiveness with categorical data. The learning rate was set to 0.0001 for the CNN and 0.001 for the BiLSTM model, maintained consistently throughout the training process. Model performance was evaluated using metrics such as accuracy, Equal Error Rate (EER). To facilitate an in-depth comparison of the models' effectiveness, the results were presented in a tabular format, enabling a clear and concise comparison across different performance metrics.

4.3 Performance of Concatenated Features

We conducted four distinct experiments leveraging two datasets, Fake or Real (FoR) and ASVSpoof 2019, employing two different combinations of concatenated feature sets: one combining MFCC, CQT, and CQCC (Feature Set 1), and another incorporating MFCC, Mel spectrograms, CQT, and CQCC (Feature Set 2). From Table 2, we observe that both concatenated feature sets, offer significant advantages in the detection of audio deepfakes, each contributing to the robustness of our approach. The first set combines the strengths of MFCC, CQT, and CQCC to provide a comprehensive spectral and cepstral analysis, achieving accuracy rates of 96.1% on the FoR dataset and 95.50% on the ASVSpoof 2019 dataset. However, the inclusion of Mel spectrograms in the second set introduces an additional layer of time-frequency information, enriching the dataset with more detailed insights into audio signals. This enhancement allows for a more nuanced detection of audio deepfakes, resulting in a testing accuracy of 97.82% on the FoR dataset and 96.63% on the ASVSpoof 2019. The superior performance of the second set is attributed to the Mel spectrograms' capability to capture richer spectral properties and subtle temporal anomalies, alongside the complementary integration of features that collectively provide a more accurate and sensitive detection mechanism against sophisticated deepfake techniques.

Table 2: Performance of Concatenated Feature Sets

Dataset	Feature Set 1		Feature Set 2	
	Accuracy	EER	Accuracy	EER
FoR	96.1%	0.042%	97.82%	0.030%
ASVSpoof2019 (LA)	95.50%	0.091%	96.63%	0.074%

The model showed a marginal improvement in performance, by about 1%, when trained on the FoR dataset compared to the ASVSpoof 2019 dataset. This improvement is attributed to the FoR dataset's broader variety of audio deepfake samples, which likely contributed to a more effective training process, enhancing the model's accuracy in distinguishing between real and fake audio.

4.4 Benchmarking

In the benchmarking analysis detailed in Table 3, our approach demonstrates promising results in audio deepfake detection, achieving accuracies of 97.82% and 96.63% on the FoR and ASVSpoof 2019 datasets, respectively. It surpasses the most comparable research by up to 3.35% in accuracy and achieves a lower EER of 0.030%. Notably, for the ASVSpoof 2019 dataset, where some studies report only EER, our method shows a significant improvement with an EER of 0.074%, compared to the 1.40% [18] and 2.82% [22] EERs reported in these studies. Such results highlight the efficacy of concatenated feature sets and the use of a BiLSTM network, setting a new benchmark in the field.

5 CONCLUSION

In this paper, we introduced a novel approach by leveraging concatenated feature sets including MFCC, Mel spectrograms, CQT, and CQCC, processed through CNN for spatial feature extraction and classified using BiLSTM networks. The proposed approach was tested across two significant datasets, the Fake or Real (FoR) and ASVSpoof 2019, demonstrating superior performance with accuracies of 97.82% and 96.63% respectively, significantly outperforming existing state-of-the-art methods. The inclusion of Mel spectrograms alongside MFCC, CQT, and CQCC in the concatenated feature set has been identified as a key factor in the enhanced

Table 3: Comparison with the state-of-the-art (the - represents a value which is not reported in the original paper)

Study	Dataset	Features	Classifier	Accuracy	EER%
[7]	FoR	MFCC, LFCC, CromaSTFT	CNN	94.47%	0.07%
[18]	ASVSpoof 2019 (LA)	CQCC, LFCC, Spec	Densely connected CNN	-	1.40%
[22]	ASVSpoof 2019 (LA)	CQCC, MFCC, LFCC, Face	DenseNet	-	2.82%
Proposed Approach	FoR	MFCC, Mel spectrogram, CQT, CQCC	BiLSTM	**97.82%**	**0.030%**
Proposed Approach	ASVSpoof 2019 (LA)	MFCC, Mel spectrogram, CQT, CQCC	BiLSTM	**96.63%**	**0.074%**

detection accuracy, providing a rich, detailed representation of audio signals that facilitates a more enhanced differentiation between real and fake audio. Future research could delve into advanced deep learning techniques for improved feature extraction, explore unsupervised learning for better handling sparse labeled data, and to enhance model robustness and generalizability.

ACKNOWLEDGMENTS

This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU and Sapienza University of Rome project 2022–2024 "EV2" (003 009 22).

REFERENCES

[1] Jahangir Alam and Patrick Kenny. 2017. Spoofing detection employing infinite impulse response—constant q transform-based feature representations. In *2017 25Th european signal processing conference (EUSIPCO)*. IEEE, 101–105.

[2] Dora M Ballesteros, Yohanna Rodriguez, and Diego Renza. 2020. A dataset of histograms of original and fake voice recordings (h-voice). *Data in brief*, 29.

[3] Nidhi Chakravarty and Mohit Dua. 2024. A lightweight feature extraction technique for deepfake audio detection. *Multimedia Tools and Applications*, 1–25.

[4] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, and Rouba Borghol. 2022. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10, 134018–134028.

[5] Lian Huang and Jinhong Zhao. 2021. Audio replay spoofing attack detection using deep learning feature and long-short-term memory recurrent neural network. In *AIIPCC 2021; The Second International Conference on Artificial Intelligence, Information Processing and Cloud Computing*. VDE, 1–5.

[6] Madeeha B Khan, Sanjay Goel, Jaswant Katar Anandan, Jersey Zhao, and Ramavath Rakesh Naik. 2022. Deepfake audio detection.

[7] Karthik Sivarama Krishnan and Koushik Sivarama Krishnan. 2023. Mfaan: unveiling audio deepfakes with a multi-feature authenticity network. *arXiv preprint arXiv:2311.03509*.

[8] Mohammed Lataifeh, Ashraf Elnagar, Ismail Shahin, and Ali Bou Nassif. 2020. Arabic audio clips: identification and discrimination of authentic cantillations from imitations. *Neurocomputing*, 418, 162–177.

[9] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2022. A comparative study on physical and perceptual features for deepfake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 35–41.

[10] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2023. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53, 4, 3974–4026.

[11] Raphaël Millière. 2022. Deep learning and synthetic media. *Synthese*, 200, 3, 231.

[12] Aakshi Mittal and Mohit Dua. 2022. Static–dynamic features and hybrid deep learning models based spoof detection system for asv. *Complex & Intelligent Systems*, 8, 2, 1153–1166.

[13] Abu Qais, Akshar Rastogi, Akash Saxena, Arpit Rana, and Deependra Sinha. 2022. Deepfake audio detection with neural networks using audio features. In *2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP)*. IEEE, 1–6.

[14] Ricardo Reimao and Vassilios Tzerpos. 2019. For: a dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 1–10.

[15] Souvik Sinha, Spandan Dey, and Goutam Saha. 2024. Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion. *Computer Speech & Language*, 86, 101599.

[16] Massimiliano Todisco et al. 2019. Asvspoof 2019: future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

[17] Xin Wang et al. 2020. Asvspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64, 101114.

[18] Zheng Wang, Sanshuai Cui, Xiangui Kang, Wei Sun, and Zhonghua Li. 2020. Densely connected convolutional network for audio spoofing detection. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1352–1360.

[19] Taiba Majid Wani and Irene Amerini. 2023. Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In *International Conference on Image Analysis and Processing*. Springer, 156–167.

[20] RLMAPC Wijethunga, DMK Matheesha, Abdullah Al Noman, KHVTA De Silva, Muditha Tissera, and Lakmal Rupasinghe. 2020. Deepfake audio detection: a deep learning based solution for group conversations. In *2020 2nd International Conference on Advancements in Computing (ICAC)*. Vol. 1. IEEE, 192–197.

[21] Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shegang Shao. 2022. Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 19–26.

[22] Junxiao Xue, Hao Zhou, Huawei Song, Bin Wu, and Lei Shi. 2023. Cross-modal information fusion for voice spoofing detection. *Speech Communication*, 147, 41–50.

[23] Jichen Yang, Rohan Kumar Das, and Haizhou Li. 2018. Extended constant-q cepstral coefficients for detection of spoofing attacks. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1024–1029.