Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Voice Spoofing Classification Using Residual Bidirectional Long Short Term Memory

Fatan Kasyidi[1*], Rifaz Muhammad Sukma[2], Annisa Mufidah Sopian[3], Dhika Rizki Anbiya[4]

[1,2,3]Department of Informatics, Universitas Jenderal Achmad Yani, Indonesia
[4] Pusat Riset Kecerdasan Artifisial dan Keamanan Siber, Organisasi Riset Elektronika dan Informatika, BRIN, Indonesia
[1,2,3]Jln Terusan Jend. Sudirman, Cibeber, Cimahi Selatan, Kota Cimahi, Indonesia
[4]Gedung B.J. Habibie, Jl. M.H. Thamrin No 8, Jakarta Pusat, Indonesia

ABSTRACT

Voice spoofing attacks are a major security concern for speech-based biometric systems. Detection and classification of spoofed voice are essential steps for preventing unauthorized accesses. This study proposes a novel approach to voice spoofing classification using a Residual Bidirectional Long Short Term Memory (R-BLSTM) network. The goal is to enhance the accuracy and robustness of voice spoofing detection using the power of deep learning and residual connections. The current proposed approach based on bidirectional LSTM with residual connections is designed to capture long-range dependencies and latent characteristics of speech signals. Experimental evidence that the R-BLSTM model is superior to classic ML techniques is also demonstrated by observing an accuracy of 95.6% on the ASVspoof 2019 collection. The designed system can be further utilized for enriching the security of speech-based biometrics modalities and making anti-voice spoofing attacks ineffective.

Keywords : *voice spoofing attacks; residual bidirectional long short term memory (R-BLSTM); ASVSpoof 2019; anti-voice spoofing.*

Article:
Accepted: July 19, 2025
Revised: March 13, 2025
Issued: October 30, 2025

***Correspondence Address:**
fatan.kasyidi@lecture.unjani.ac.id

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

1. INTRODUCTION

Spoofing has emerged as a significant threat that compromises the integrity and security of various systems. Spoofing refers to the malicious practice of deceiving a system by impersonating another user or device, often to gain unauthorized access to sensitive information or manipulate data. This technique can take various forms, including IP spoofing, email spoofing, and more recently, voice spoofing, in which attackers use synthetic or altered voices to deceive voice recognition systems [1], [2], [3]. Detecting and preventing spoofing are critical for maintaining the security and trustworthiness of digital systems. Robust spoofing detection algorithms are crucial for protecting confidential information, preserving user identity integrity, and maintaining the efficacy of authentication protocols in digital systems [4], [5].

As voice recognition technology has gained prominence in security systems, financial services, and personal devices, the need for robust voice spoofing detection methods has become even more crucial. Recent studies have highlighted the vulnerability of automatic speaker verification (ASV) systems to various spoofing attacks, including those utilizing synthesized speech and voice conversion techniques [6], [7], [8]. Without effective detection, systems are vulnerable to attacks that can result in significant financial losses, privacy breaches, and loss of user trust [9], [10]. In this study, we address the challenge of voice spoofing detection using the ASVSpoof 2019 dataset. A primary issue with this dataset is its imbalanced nature, in which certain classes are underrepresented [2], [11], [12]. This imbalance can lead to models that are either overfitted to the majority class or underfitted to the minority class, thereby reducing the overall effectiveness of the spoofing detection system [13], [14].

The increasing sophistication of spoofing techniques has necessitated the development of advanced detection methods that can effectively identify and mitigate these threats. The ASVSpoof initiative has been instrumental in providing datasets and benchmarks for research in this area, enabling the development of more effective antispoofing solutions [2], [13], [15]. As the landscape of voice recognition and synthesis continues to evolve, ongoing research and innovation in spoof detection will be essential for safeguarding digital interactions and maintaining user trust.

Several studies have been conducted on spoofing detection methods. Mawada et al studied about the use of inverted MFCC (iMFCC) combined with high-frequency features of spoof detection [16]. The results show that BiLSTM achieved 99.58% validation accuracy and 6.58% error rate with ASVSpoof 2017 as the dataset. Another research conducted by Chaudari and Shedge shows that using combined MFCC and CQCC the error rate for detecting voice spoofing using AVSpoof 2017 dataset is 10,18% [17]. Qadir et al. studied the use of LSTM with MFCC, GTCC, and spectral centroid as features, and achieved 1.30% EER with the ASVSpoof 2019 LA dataset [6]. Anagha et al used CNN with Mel Spectrogram images as a feature to detect voice spoofing from ASVSpoof 2019 dataset. It is shows that 85% of an accuracy[18].

Furthermore, we followed the same method for selecting the dataset used in previous research. Specifically, audio replay classes were excluded, and the first-degree and second-degree replay data were merged into a single spoof class. This standardization ensured consistency with previous studies and focused our analysis on the most relevant data. To overcome this, we propose the use of the Residual Bidirectional Long Short-Term Memory (BiLSTM) method for voice spoof classification, employing Constant Q Cepstral Coefficients (CQCC) and mel-frequency cepstral coefficients (MFCC) as features. However, based on related research, the error rate is high. We propose using a dual-layer BiLSTM to overcome this problem. Additionally, to address the data imbalance issue, we applied various balancing techniques, such as Random Sampling.

By integrating these techniques, our study aimed to provide a more accurate and reliable method for detecting voice spoofing, thereby contributing to a broader effort to enhance cybersecurity measures against such sophisticated attacks.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

2. METHODS

This work begins with dataset selection, and we used data from ASVSpoof2019. The next step is feature extraction of the signals using two different feature extraction methods, namely, Mel-Frequency Cepstral Coefficients (MFCC) and constant Q cepstral coefficients (CQCC). Furthermore, the signals processed through feature extraction are used as the input in the classification stage. The classification model used in this research is Residual BiLSTM, which classifies the signal data into two classes: Spoof and Bonafide. Figure 1 shows the workflow of this study.

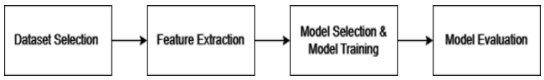

Figure 1. *Workflow*

2.1. Data Selection

At this stage, data is collected using the data In this study, we utilized the ASVSpoof 2019 dataset, which has become a cornerstone for research on automatic speaker verification (ASV). ASV is recognized as one of the most effective biometric recognition methods, leveraging unique vocal characteristics to authenticate individuals [19]. Recent advancements in ASV systems have significantly improved their performance; however, they remain susceptible to various forms of manipulation, collectively termed spoofing or presentation attacks [20].

The ASVspoof 2019 challenge specifically addresses three types of spoofing attacks, namely, speech synthesis, voice conversion, and replay attacks, all within two distinct scenarios: logical access (LA) and physical access (PA) [19], [20]. Each scenario utilized different datasets, allowing for comprehensive exploration of the vulnerabilities inherent in ASV systems. The dataset is derived from the VCTK corpus, featuring recordings from 107 speakers (46 male and 61 female). These speakers are partitioned into training, development, and evaluation sets, ensuring speaker disjointness across these partitions. The utterances are downsampled to 16 kHz to standardize the audio data. The ASVspoof 2019 dataset is divided into two scenarios, Logical Access (LA) and Physical Access (PA). The LA scenario comprises 12,483 bona fide and 108,978 spoofed utterances, totaling 121,461 samples, while the PA scenario includes 28,890 bona fide and 189,540 spoofed utterances, amounting to 218,430 samples. In total, the dataset contains 41,373 bona fide and 298,518 spoofed utterances, resulting in 339,891 samples.

In our study, we focused on a subset of 5,000 data samples from the ASVspoof 2019 dataset, specifically from the Logical Access scenario, to analyze the effectiveness of the current anti-spoofing measures.

2.2. Feature Extraction

Feature extraction is necessary for obtaining valuable characteristics from audio signals. In this study, we used two different feature extraction methods: MFCC and CQCC. The MFCC scales the frequencies to fit better with those that can be heard by humans, as the human ear perceives actual frequencies differently [21]. It is difficult for humans to distinguish between high and low frequencies. Therefore, a scale called the MEL scale is used to transform the actual frequencies into perceived frequencies. Mel-frequency cepstral coefficients (MFCCs) and their variant, the complex cepstral coefficients (CQCC), are pivotal in automatic speaker verification (ASV) systems, particularly in the context of voice signal research. MFCCs are widely recognized for their effectiveness in capturing the spectral characteristics of speech signals and in transforming audio data into a set of coefficients that represent the short-term power spectrum of sound. This transformation is crucial for distinguishing between different speakers and improving recognition accuracy in ASV systems [22], [23]. Recent studies have demonstrated that MFCCs can be particularly beneficial for processing degraded audio signals, thereby enhancing speaker recognition performance through the application of Gaussian Mixture Models (GMM) [24], [25]. CQCC, which incorporates complex spectral information, has been shown to outperform traditional MFCC in scenarios involving spoofing attacks because it effectively captures the

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

phase information that is often lost in standard cepstral analysis, which is critical for distinguishing between genuine and spoofed audio signals [26], [27], [28]. The approach of these two feature extractions can help to recognize attacks in detail [29].

2.3. Model Deep Learning

The deep learning model for voice signal attack classification in 2023 utilized a sophisticated architecture combining Recurrent Neural Networks (RNN) with Bidirectional Long Short-Term Memory (BiLSTM) layers[30]. This model is structured with two initial RNN layers employing the BiLSTM architecture, each consisting of 64 filters, which effectively capture temporal enhances the ability to capture temporal dependencies in audio signals effectively [31], [32]. The incorporation of residual connection networks in these layers facilitates improved gradient flow during training, thereby addressing the vanishing gradient problem commonly encountered in deep networks [33], [34]. Following the RNN layers, the output data were flattened into vectors and fed into a Fully Connected (FC) layer comprising three dense networks. The first two dense layers incorporate dropout with a rate of 0.5 and Batch Normalization, utilizing ReLU activation functions to mitigate overfitting and enhance convergence during training [35], [36]. The final dense layer employs a sigmoid activation function, which is particularly suited for binary classification tasks, ensuring that the model outputs probabilities indicative of the presence of voice signal attacks [37], [38]. The architecture in this research exemplifies the integration of advanced deep learning techniques to improve the robustness and accuracy of speech-signal classification systems. An illustration of the designed architectural model is shown in figure 2.

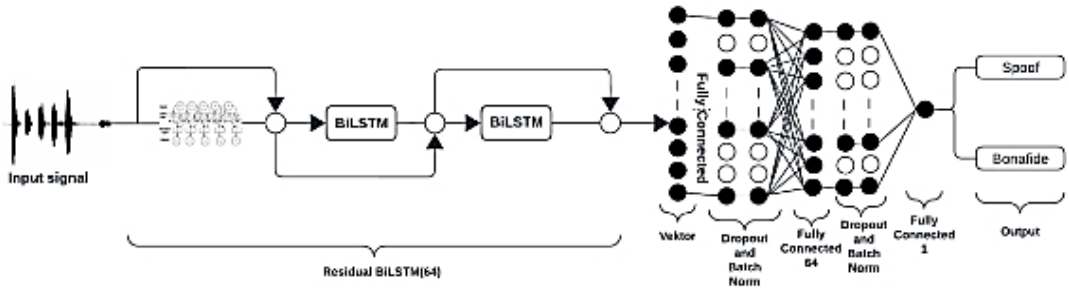

Figure 2. *Illustration of the deep learning model used, which uses a combination of rnn and bilstm architectures. the first and second rnn layers use the bilstm architecture with 64 filters and a residual connection network; then, the data are flattened into vectors to enter fully connected with three dense networks, where the first and second dense networks add a dropout of 0.5, batch normalization with relu activation, and the last dense network uses sigmoid activation*

2.4. Model Training

In this study, the training configuration for the deep learning model employed in voice signal attack classification was meticulously designed to optimize performance and accuracy. Epochs were set to a number of 50 with a batch size of 32. The learning rate is set to 0.0001, which is a common choice that balances the speed of convergence with the stability of the training process, allowing the model to learn effectively without overshooting the optimal solution [39]. The Adam optimizer is utilized, known for its adaptive learning rate capabilities, which helps in efficiently navigating the loss landscape, particularly in complex models[40]. The loss function chosen is binary cross-entropy, appropriate for binary classification tasks, as it quantifies the difference between the predicted probabilities and the actual class labels, thereby guiding the model towards minimizing classification errors [41]. This comprehensive training configuration underscores the importance of hyperparameter tuning in achieving optimal model performance in deep learning tasks.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

2.5. Model Evlauation

In this study, the evaluation of a deep learning model for voice signal attack classification was conducted using several key performance metrics, including accuracy, precision, recall, F1 score, and Equal Error Rate (EER). Precision measures the accuracy of the model's positive predictions, indicating how many of the predicted positive instances are actually true positives, while recall assesses the model's ability to identify all relevant instances, reflecting the proportion of true positives out of the total actual positives [42]. The F1 score serves as a harmonic mean of precision and recall, providing a single metric that balances both aspects, and is particularly useful in scenarios with imbalanced datasets [43]. Furthermore, the Equal Error Rate (EER) is employed as a critical metric in binary classification tasks, representing the point at which the false positive rate equals the false negative rate, thus offering a comprehensive view of the model's performance across different thresholds [44]. This multifaceted evaluation approach ensures a robust assessment of the effectiveness of the model in accurately classifying voice-signal attacks, highlighting its strengths and areas for improvement. Together, these metrics provide a robust framework for evaluating the effectiveness and reliability of deep learning models in classifying voice-signal attacks.

$$EER = (FAR + FRR)/2 \quad (1)$$

$$ACC = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

$$Precission = \frac{TP}{(TP+FP)} \quad (3)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

$$F1\ Score = \frac{2*(Recall*Precission)}{(Recall*Precission)} \quad (5)$$

3. RESULTS AND DISCUSSION

The exploration of Constant-Q Cepstral Coefficients (CQCC) as a feature extraction method for spoofing detection in Automatic Speaker Verification (ASV) systems has shown promising results, however, there is still a research gap, especially with regard to the LA 2019 spoof ASV dataset [45]. Previous studies have mostly focused on the efficacy of CQCC in various contexts, such as replay attack detection and hypernasality severity detection, but they often ignore the unique challenges posed by the LA 2019 spoof dataset, which includes various spoofing techniques and environmental conditions [46], [47]. Although CQCC has been established as a robust feature extraction method, its applicability has not been comprehensively evaluated against the unique characteristics of the LA 2019 spoof dataset, which may lead to poor performance in real-world scenarios [48], [49]. Furthermore, existing studies mainly use traditional machine learning classifiers, such as Gaussian Mixture Models (GMMs), without utilizing advanced deep learning architectures that have the potential to improve feature representation and classification accuracy [50], [51]. This highlights the critical need to develop novel deep learning models that integrate CQCC features specifically tailored to the complexity of the LA 2019 spoof dataset, thereby overcoming the limitations of previous methodologies and improving the robustness of spoof detection systems [52], [53].

This study aims to develop a deep learning model for spoofing classification by utilizing the ASVspoof 2019 LA dataset, specifically focusing on the extraction of features using Constant-Q Cepstral Coefficients (CQCC) and comparing their effectiveness against Mel-Frequency Cepstral Coefficients (MFCC). The ASVspoof 2019 dataset includes a diverse range of spoofing attacks such as speech synthesis and voice conversion, providing a robust framework for evaluating the performance of different feature extraction techniques in the context of automatic speaker verification (ASV) [54], [55].

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

Previous studies have highlighted the advantages of CQCC in capturing perceptually relevant features, which may enhance the detection of sophisticated spoofing attacks compared to traditional MFCC methods[46]. Moreover, the integration of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has shown promise in improving the classification accuracy by leveraging the temporal dependencies in audio signals [56]. By systematically comparing the performance of CQCC and MFCC features within a deep learning framework, this research seeks to contribute to the ongoing efforts to enhance the robustness of ASV systems against logical access attacks [56], [57].

In this study, the spoof detector utilizes the ASVspoof2019 dataset, which was compiled in a balanced manner. The distribution of the data used is shown in Table (1), which helps ensure a balanced representation of different types of spoofing. This shows that the research was conducted with respect to the diversity and consistency of the data from the dataset.

Table 1. *Data distribution*

Labels	Train		Test	
	CQCC	MFCC	CQCC	MFCC
Spoof	1822	1822	790	790
Bonafide	1790	1790	758	758

After data collection, the next stage involved feature extraction from the sound waves using the CQCC and MFCC feature extraction methods. This process aims to extract important information from sound signals that can be used for further analysis. By using the CQCC and MFCC feature methods, a robust representation of the main features of each generated sound sample is expected, as shown in figure (3). The results of sound wave feature extraction show that CQCC feature extraction is more effective in detailing the characteristics of spoof sounds than the MFCC method. The use of CQCC provides the ability to capture spectral information better, especially at high frequencies, which is often an important area in detecting voice manipulation. This results in a richer and more accurate feature representation for distinguishing between the original and spoof voices. Utilizing the CQCC features.

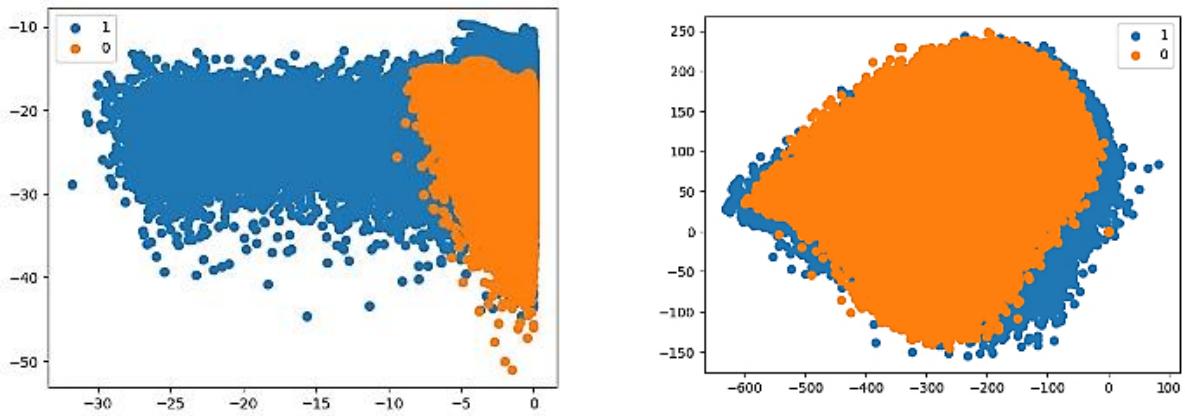

Figure 3. *Visualization and comparison between CQCC and MFCC feature extraction, where the left is the result of CQCC feature extraction, which shows how detailed the resulting features are, inversely proportional to MFCC feature extraction, which has a similarity between the two labels on the left*

After obtaining the important features that represent the characteristics of the spoof and bona fide voices, the next step is to train the feature results using deep learning. The deep learning model is carefully constructed in figure (2) to approach the level of perfection, so as to be able to distinguish between spoof and bona fide voices with high accuracy. This training process aims to enable the model to recognize the complex and abstract patterns contained in voice data so that it can perform classification with maximum precision.

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (184-195)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.43281

The results of model training using a model architecture that is a combination of RNNs, particularly BiLSTM, show satisfactory performance. This model architecture consists of the first and second three RNN layers, each of which uses the BiLSTM architecture with 64 filters and a residual connection network. Subsequently, the data are flattened into vectors to feed into a fully connected layer consisting of three dense layers. In the first and second dense layers, a Dropout of 0.5 and Batch Normalization with Relu activation were applied, while the last dense layer used sigmoid activation.

After training the model with this dataset, it showed good accuracy; the training parameters are listed in Table (2). Based on the observations, it can be seen in Figure (4) that the accuracy and loss values for the training data and validation data begin to stabilize and converge at the 40th to 50th epoch, especially in the third graph. This shows that the model does not experience overfitting, as the accuracy and loss curves between the training and validation data remain in line without any significant difference. With this consistency in performance, the model demonstrates a good ability to distinguish between real and spoof voices, indicating that the model architecture has been effectively designed to cope with the complexity and variation in voice data.

Table 2. *Configuration of parameters training*

Parameters	Value
epoch	50
batch_size	32
learning_rate	0.0001
optimize	Adam
loss	binary_crossentropy
metrics	accuracy

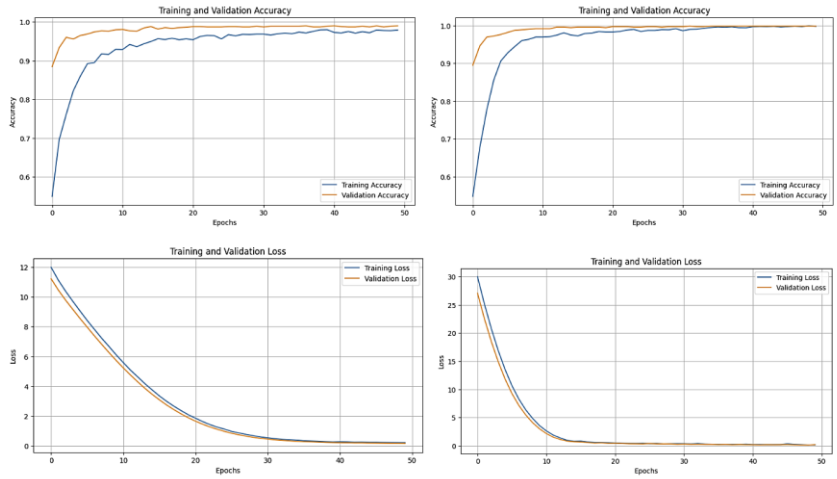

Figure 4. *Visualization and comparison between accuracy and loss on training model with CQCC and MFCC features, where the left is the result of CQCC feature extraction and the right with MFCC feature extraction*

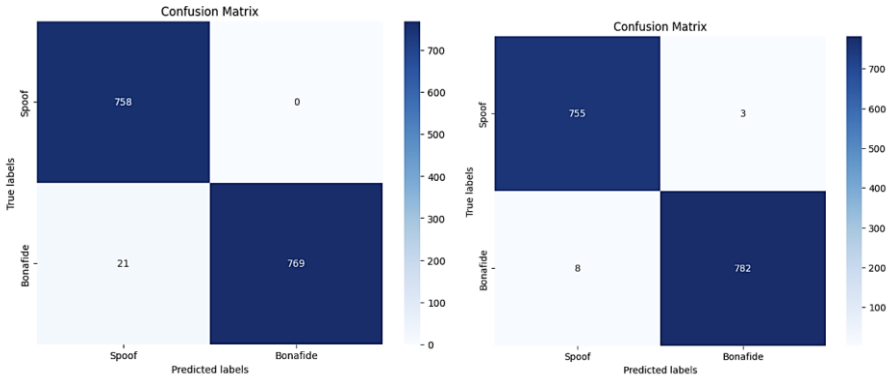

Figure 5. *Visualization and comparison between CQCC and MFCC feature extraction, where the left is the result of CQCC feature extraction, which shows how detailed the resulting features are, inversely proportional to MFCC feature extraction, which has a similarity between the two labels on the left*

The model evaluation showed good performance in classifying the Spoof and Bonafide labels. For the CQCC feature, the Precision, Recall, and F1-Score values for the Spoof label are 97.3%, 100%, and 98.63%, respectively, while for the Bonafide label, it has a precision of 100%, recall of 97.34%, and F1-Score of 98.65%. These results indicate that the model was able to detect both labels with fairly high accuracy, with an Error Equal Rate (EER) of 0.013, indicating a low misclassification rate. In contrast, MFCC performed better than CQCC. The MFCC recorded Precision, Recall, and F1-Score for Spoof labels of 98.95%, 99.6%, and 99.28%, respectively, while Bonafide labels had precision of 99.62%, recall of 98.99%, and F1-Score of 99.3%. The lower EER rate of MFCC (0.007) shows that this feature is more effective in reducing misclassification. However, the model created with CQCC feature extraction showed excellent performance in detecting Spoof labels, as evidenced by the achievement of a recall value of 100%. This means that the model successfully recognizes all spoof samples without missing any samples. This result indicates the reliability of the CQCC feature in capturing the distinctive characteristics of the Spoof data so that all samples with such labels can be perfectly identified.

Table 3. Result of evaluation model.

Feature Extraction	Labels	Precision	Recall	F1-Score	Accuracy	Error Equal Rates
CQCC	Spoof	97.3%	100%	98.63%	98.64%	0.013
	Bonafide	100%	97.34%	98.65%		
MFCC	Spoof	98.95%	99.6%	99.28%	99.29%	0.007
	Bonafide	99.62%	98.99%	99.3%		

The results show that the deep learning model developed in this study performs well, with an improvement in the Equal Error Rate (EER) value. The proposed model achieves an EER of 1.30% and 0.70%, which is lower than the previous study that using LSTM with MFCC, GTCC, and spectral centroid recorded an EER of 1.30% [6]. This decrease in EER indicates an increase in the model's ability to detect voice forgery more accurately.

For future research, it is recommended that this model be tested with a different dataset, for example, the ASVspoof 2021 dataset. ASVspoof 2021 has some significant differences compared to ASVspoof 2019, particularly in terms of task scope and challenge complexity. ASVspoof 2021 introduces a new task that focuses on speech forgery detection, which expands the scope beyond logical and physical access, as was the main focus of ASVspoof 2019. In addition, ASVspoof 2021 includes more realistic scenarios with new trials that account for voice encoding and transmission artifacts, reflecting situations where voices are transmitted through various phone systems[58] [59]. Given these additional challenges, ASVspoof 2021 is expected to provide a more rigorous test for the developed voice-forgery models.

CONCLUSION

The results showed that the extraction of CQCC features with the deep learning models used in this training obtained excellent results in detecting spoofing with a recall value of 100%. This shows that this combination is suitable for spoofing detection; however, there are times when Bonafide is identified as a spoof, with an ERR evaluation of 1.30%. In addition, the combination of MFCC and deep learning models showed better ERR results (0.70%); however, the recall on spoofing was only 99.6%. To see the results of more in-depth research, further research using a combination of MFCC and deep learning models is required. In addition, the combination of MFCC and the deep learning model showed better ERR results (0.70 %); however, the recall on spoof was only 99.6%. To see the results of more in-depth research, further research is needed using the combination of this study with different datasets, namely the ASVspoof 2021 dataset.

REFERENCES

[1] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in Anti-Spoofing: From the Perspective of ASVspoof Challenges," *APSIPA Trans Signal Inf Process*, 2020, doi: 10.1017/atsip.2019.21.

[2] A. Nautsch *et al.*, "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Trans Biom Behav Identity Sci*, 2021, doi: 10.1109/tbiom.2021.3059479.

[3] A. Kuznetsov, R. A. Murtazin, I. M. Garipov, E. A. Fedorov, A. V Kholodenina, and A. Vorobeva, "Methods of Countering Speech Synthesis Attacks on Voice Biometric Systems in Banking," *Scientific and Technical Journal of Information Technologies Mechanics and Optics*, 2021, doi: 10.17586/2226-1494-2021-21-1-109-117.

[4] C.-W. Bang, "Effective Zero-Shot Multi-Speaker Text-to-Speech Technique Using Information Perturbation and a Speaker Encoder," *Sensors*, 2023, doi: 10.3390/s23239591.

[5] J. Guo, Z. Yun-yu, and H. Wang, "Generalized Spoof Detection and Incremental Algorithm Recognition for Voice Spoofing," *Applied Sciences*, 2023, doi: 10.3390/app13137773.

[6] G. Qadir, S. Zareen, F. Hassan, and A. U. Rahman, "Voice Spoofing Countermeasure Based on Spectral Features to Detect Synthetic Attacks Through LSTM," *International Journal of Innovations in Science and Technology*, vol. 3, no. 5, pp. 153–165, Jan. 2022, doi: 10.33411/ijist/2021030512.

[7] T. Kaichi and Y. Ozasa, "A Hyperspectral Approach for Unsupervised Spoof Detection With Intra-Sample Distribution," 2021, doi: 10.1109/icip42928.2021.9506625.

[8] H. Tak, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," 2022, doi: 10.48550/arxiv.2202.12233.

[9] Y. Zhang, "One-Class Learning Towards Synthetic Voice Spoofing Detection," 2020, doi: 10.48550/arxiv.2010.13995.

[10] I. Gurowiec, "Speech Emotion Recognition Systems and Their Security Aspects," *Artif Intell Rev*, 2024, doi: 10.1007/s10462-024-10760-z.

[11] E. Wenger *et al.*, "'Hello, It's Me': Deep Learning-Based Speech Synthesis Attacks in the Real World," 2021, doi: 10.1145/3460120.3484742.

[12] C. Hu and R. Zhou, "Synthetic Voice Spoofing Detection Based on Online Hard Example Mining," 2022, doi: 10.22541/au.166429442.20902648/v1.

[13] J. Guo and Z. Zhao, "Generalized Spoof Detection and Incremental Algorithm Recognition for Spoofing Voice," 2022, doi: 10.21203/rs.3.rs-2149586/v1.

[14] H. Zeinali *et al.*, "Detecting Spoofing Attacks Using VGG and SincNet: BUT-Omilia Submission to ASVspoof 2019 Challenge," 2019, doi: 10.21437/interspeech.2019-2892.

[15] A. Chadha, A. Abdullah, and L. Angeline, "A Unique Glottal Flow Parameters Based Features for Anti-Spoofing Countermeasures in Automatic Speaker Verification," *International Journal of Advanced Computer Science and Applications*, 2021, doi: 10.14569/ijacsa.2021.0120894.

[16] H. Mewada *et al.*, "Gaussian-Filtered High-Frequency-Feature Trained Optimized BiLSTM Network for Spoofed-Speech Classification," *Sensors*, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146637.

[17] A. Chaudhari and D. K. Shedge, "Integration of CQCC and MFCC based Features for Replay Attack Detection," in *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2022, pp. 1–5. doi: 10.1109/ESCI53509.2022.9758391.

[18] R. Anagha, A. Arya, V. H. Narayan, S. Abhishek, and T. Anjali, "Audio Deepfake Detection Using Deep Learning," in *Proceedings of the 2023 12th International Conference on System Modeling and Advancement in Research Trends, SMART 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 176–181. doi: 10.1109/SMART59791.2023.10428163.

[19] M. Todisco *et al.*, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," 2019, doi: 10.21437/interspeech.2019-2249.

[20] C. B. Tan *et al.*, "A Survey on Presentation Attack Detection for Automatic Speaker Verification Systems: State-of-the-Art, Taxonomy, Issues and Future Direction," *Multimed Tools Appl*, 2021, doi: 10.1007/s11042-021-11235-x.

[21] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma," Nov. 2020, pp. 87–91. doi: 10.1109/CICN49253.2020.9242635.

[22] U. Ayvaz, H. Gürüler, F. U. Khan, N. Ahmed, T. K. Whangbo, and A. A. Bobomirzaevich, "Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning," *Computers Materials & Continua*, 2022, doi: 10.32604/cmc.2022.023278.

[23] M. Neelima and I. S. Prabha, "Spoofing Detection and Countermeasure in Automatic Speaker Verification System Using Dynamic Features," *International Journal of Recent Technology and Engineering*, 2020, doi: 10.35940/ijrte.e6582.018520.

[24] A. Moondra and P. Chahal, "Improved Speaker Recognition for Degraded Human Voice Using Modified-MFCC and LPC With CNN," *International Journal of Advanced Computer Science and Applications*, 2023, doi: 10.14569/ijacsa.2023.0140416.

[25] A. Moondra and P. Chahal, "Speaker Recognition Improvement for Degraded Human Voice Using Modified-MFCC With GMM," *International Journal of Advanced Computer Science and Applications*, 2023, doi: 10.14569/ijacsa.2023.0140627.

[26] B. Bhagat and M. Dua, "Enhancing Performance of End-to-End Gujarati Language ASR Using Combination of Integrated Feature Extraction and Improved Spell Corrector Algorithm," *Itm Web of Conferences*, 2023, doi: 10.1051/itmconf/20235401016.

[27] K. Phapatanaburi, L. Wang, S. Nakagawa, and M. Iwahashi, "Replay Attack Detection Using Linear Prediction Analysis-Based Relative Phase Features," *Ieee Access*, 2019, doi: 10.1109/access.2019.2960369.

[28] D. Li *et al.*, "Multiple phase information combination for replay attacks detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2018, pp. 656–660. doi: 10.21437/Interspeech.2018-2001.

[29] R. Jahangir *et al.*, "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network," *Ieee Access*, 2020, doi: 10.1109/access.2020.2973541.

[30] S. Ibrar, A. Javed, and H. Ilyas, "Voice Presentation Attacks Detection using Acoustic MLTP Features and BiLSTM," in *2023 3rd International Conference on Communication, Computing and Digital Systems, C-CODE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/C-CODE58145.2023.10139903.

[31] C. Wall, L. Zhang, Y. Yu, and K. Mistry, "Deep Recurrent Neural Networks With Attention Mechanisms for Respiratory Anomaly Classification," 2021, doi: 10.1109/ijcnn52387.2021.9533966.

[32] W. Huang, X. Liu, M. Luo, P. Zhang, W. Wang, and J. Wang, "Video-Based Abnormal Driving Behavior Detection via Deep Learning Fusions," *Ieee*

Access, 2019, doi: 10.1109/access.2019.2917213.

[33] Y. R. Musunuri and O. Kwon, "Deep Residual Dense Network for Single Image Super-Resolution," *Electronics (Basel)*, 2021, doi: 10.3390/electronics10050555.

[34] Y. Han, P. Cui, Y. Zhang, R.-G. Zhou, S. Yang, and J. Wang, "Remote Sensing Sea Ice Image Classification Based on Multilevel Feature Fusion and Residual Network," *Math Probl Eng*, 2021, doi: 10.1155/2021/9928351.

[35] D. Chen, F. Hu, G. Nian, and T. Yang, "Deep Residual Learning for Nonlinear Regression," *Entropy*, 2020, doi: 10.3390/e22020193.

[36] B.-C. Yang and G. Wu, "Efficient Single Image Super-Resolution Using Dual Path Connections With Multiple Scale Learning," 2021, doi: 10.48550/arxiv.2112.15386.

[37] B. Liu, K. Gao, A. Yu, W. Guo, R. Wang, and X. Zuo, "Semisupervised Graph Convolutional Network for Hyperspectral Image Classification," *J Appl Remote Sens*, 2020, doi: 10.1117/1.jrs.14.026516.

[38] J. Wu, W. Hu, Y. Wen, W. Tu, and X. Liu, "Skin Lesion Classification Using Densely Connected Convolutional Networks With Attention Residual Learning," *Sensors*, 2020, doi: 10.3390/s20247080.

[39] Y. Zhang, "Deep Learning Distributed Architecture Design Implementation for Computer Vision," *Wirel Commun Mob Comput*, 2022, doi: 10.1155/2022/9726286.

[40] I. A. Klampanos, A. Davvetas, A. Koukourikos, and V. Karkaletsis, "ANNETT-O: An Ontology for Describing Artificial Neural Network Evaluation, Topology and Training," *Int J Metadata Semant Ontol*, 2019, doi: 10.1504/ijmso.2019.099833.

[41] Y.-K. Lee, W. Sim, P. Jeongmook, and J. Lee, "Evaluation of Hyperparameter Combinations of the U-Net Model for Land Cover Classification," *Forests*, 2022, doi: 10.3390/f13111813.

[42] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis Using Various Performance Metrics in Imbalanced Data for Multi-Class Text Classification," *International Journal of Advanced Computer Science and Applications*, 2023, doi: 10.14569/ijacsa.2023.01406116.

[43] A. A. Alnuaim *et al.*, "Speaker Gender Recognition Based on Deep Neural Networks and ResNet50," *Wirel Commun Mob Comput*, 2022, doi: 10.1155/2022/4444388.

[44] S. Sen, S. Maiti, S. Manna, B. Roy, and A. Gosh, "Smart Prediction of Water Quality System for Aquaculture Using Machine Learning Algorithms," 2023. doi: 10.36227/techrxiv.22300435.v1.

[45] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," *Machine Learning with Applications*, vol. 14, p. 100503, Dec. 2023, doi: 10.1016/j.mlwa.2023.100503.

[46] A. Mittal and M. Dua, "Static–dynamic Features and Hybrid Deep Learning Models Based Spoof Detection System for ASV," *Complex & Intelligent Systems*, 2021, doi: 10.1007/s40747-021-00565-w.

[47] M. Adiban, H. Sameti, and S. Shehnepoor, "Replay Spoofing Countermeasure Using Autoencoder and Siamese Network on ASVspoof 2019 Challenge," 2019, doi: 10.48550/arxiv.1910.13345.

[48] C. Hu, "Synthetic Speech Spoofing Detection Based on Online Hard Example Mining," *Ieee Access*, 2023, doi: 10.1109/access.2023.3311849.

[49] X. Cheng, M. Xu, and T. F. Zheng, "A Multi-Branch ResNet With Discriminative Features for Detection of Replay Speech Signals," *APSIPA Trans Signal Inf Process*, 2020, doi: 10.1017/atsip.2020.26.

[50] H. Yu, Z. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features," *IEEE Trans Neural Netw Learn Syst*, 2018, doi: 10.1109/tnnls.2017.2771947.

[51] Q. Wang, K. A. Lee, and T. Koshinaka, "Using Multi-Resolution Feature Maps With Convolutional Neural Networks for Anti-Spoofing in ASV," 2020, doi: 10.21437/odyssey.2020-20.

[52] J. Li, M. Sun, X. Zhang, and Y. Wang, "Joint Decision of Anti-Spoofing and Automatic Speaker Verification by Multi-Task Learning With Contrastive Loss," *Ieee Access*, 2020, doi: 10.1109/access.2020.2964048.

[53] A. Kumar, D. Paul, M. Pal, M. Sahidullah, and G. Saha, "Speech Frame Selection for Spoofing Detection With an Application to Partially Spoofed Audio-Data," *Int J Speech Technol*, 2021, doi: 10.1007/s10772-020-09785-w.

[54] L. Wei, Y. Long, H. Wei, and Y. Li, "New Acoustic Features for Synthetic and Replay Spoofing Attack Detection," *Symmetry (Basel)*, 2022, doi: 10.3390/sym14020274.

[55] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU Replay Detection System for the ASVspoof 2019 Challenge: On Data Augmentation, Feature Representation, Classification, and Fusion," 2019, doi: 10.21437/interspeech.2019-1230.

[56] T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice Spoofing Countermeasure for Logical Access Attacks Detection," *Ieee Access*, 2021, doi: 10.1109/access.2021.3133134.

[57] Y. Zhao, R. Togneri, and V. Sreeram, "Multi-Task Learning-Based Spoofing-Robust Automatic Speaker Verification System," *Circuits Syst Signal Process*, 2022, doi: 10.1007/s00034-022-01974-z.

[58] X. Dang, Z. Zhao, and N. Wu, "Research on Speech Playback Spoof Detection Based on ASV Spoof 2021," in *Proceedings of 2024 International Conference on New Trends in Computational Intelligence, NTCI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 538–543. doi: 10.1109/NTCI64025.2024.10776128.

[59] J. Yamagishi *et al.*, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," no. September, pp. 47–54, 2021, doi: 10.21437/asvspoof.2021-8.