ISSN: 2093-5374 / E-ISSN: 2093-5382

ResNet152: A Deep Learning Approach for Robust Spoof Detection in Speaker Verification Systems

M. Selin[1*], and Dr.K. Preetha Mathew[2]

[1*]Department of Computer Applications, Cochin University of Science and Technology, Cochin, Kerala, India. selin.m.a@gmail.com, https://orcid.org/0000-0002-9404-4092

[2]Cochin University College of Engineering, Kuttanad, Pulincunnu, Alappuzha, Kerala, India. preetha.mathew.k@gmail.com, https://orcid.org/0009-0000-1870-3618

Received: June 15, 2024; Revised: August 04, 2024; Accepted: August 30, 2024; Published: September 30, 2024

Abstract

In human life, we know that sound is the most important factor. From the normal perspective to the intelligent perspective, sound develops automated systems for various fields for several purposes. However, within contemporary conventional systems, there is significant abuse leading to the proliferation of forgery and other crimes, with sound often playing a central role. With the help of the latest technology such as deep learning, there comes a vast possibility of integrating with many systems for boosting the efficiency of existing systems. So, in this paper, we bring an effective classification of audio using ResNet152. The audio signals are converted to spectrogram images and are passed to a classifier for generating binary classification such as genuine or spoof. We also evaluated our model with existing methods such as VGG16, CNN, VGG19, and AlexNet under performance measures such as Accuracy, EER, and t-DCF in which the proposed model outperforms with 92.2% testing accuracy and 82.2% inference accuracy.

Keywords: Audio Spoof, ASVSpoof2019, Classification, Deep Learning, ResNet152.

1 Introduction

Biometrics technology has become increasingly popular as a result of the Internet's rapid expansion and is being used extensively in a variety of disciplines, including medical education, financial and social security, criminal investigation for public safety, intelligent security, and criminal justice (Boulkenafet et al., 2015; Srinivasa Rao et al., 2023). Voice recognition technology is becoming a major area of research for both academia and industry because of its benefits over existing biometric recognition technology, including safety, naturalness, and non-contact. Nonetheless, the speech recognition system's security performance is seriously threatened by harmful assaults by unauthorized users (Verkholyak et al., 2021). As a result, creating an anti-spoofing system with great durability, quick response times, and high detection accuracy is crucial.

Also, with the advancement of Deep Learning (DL) technology we have been using this over so many fields mostly in this audio spoofing field, DL has revolutionized better than expected. Various researchers have built various DL models for audio spoofing detection with the priority of improving the accurate results in predicting voices that are genuine or spoofed (Udayakumar et al., 2023).

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 3 (September), pp. 354-363. DOI: *10.58346/JOWUA.2024.I3.023*

*Corresponding author: Department of Computer Applications, Cochin University of Science and Technology, Cochin, Kerala, India.

Countermeasures that are currently in place and aim to identify specific spoofing attempts usually rely on past knowledge about a given spoofing algorithm (Kinnunen et al., 2012). Consequently, these remedies cannot be used for different types of spearing assaults (Kinnunen et al., 2017). When comparing spoof recordings to authentic recordings, one usually searches for the audio features that have been altered the most or the least during the parametrization process. To efficiently model the acquired audio data and reliably anticipate speech, different modelling approaches are frequently evaluated at the back end (Gümüş et al., 2022).

Key Highlights

Here are some highlights of this paper, which focuses on creating an efficient deep-learning model for audio spoof detection:

a. Audio spoof detection using the ResNet152 model.
b. Pre-processing these audio files and converting them to spectrogram for effective analysis.
c. With the help of a classifier, we can generate a binary classification like genuine or spoof audio.
d. Evaluating our model with other existing methods in which the proposed model gains better accuracy.

Organization of Paper: Since we have already read the introduction in Section 1, the remaining sections are as follows: Section 2 lists relevant works, Section 3 explains the framework's approach, Section 4 presents the performance evaluation, and Section 5 presents the conclusion.

2 Related Works

Table 1 presents a comprehensive review of different deep-learning models for spoof detection in speaker verification systems in various ASVspoof datasets (Wong & Yiu, 2020).

Table 1: summarizes a Review of Various Deep-Learning Techniques for Spoof Detection in Automatic Speaker Verification

Cited	Dataset	Features Used	Classification Model	EER(%)
(Hanilçi et al., 2015)	ASVspoof 2015	MFCC	GMM-ML (Maximum likelyhood)	3.01%
(Pal et al., 2018)	ASVspoof 2015	CQCC, APGDF, Fundamental Frequency Variation	GMM (Gaussian Mixture Model)	0.05%
(Scardapane et al., 2017)	ASVspoof 2015	MFCC	Deep RNN (Recurrent Neural Network)	2.910%
(Zhao et al., 2018)	ASVspoof 2015	CQCC, SCC	GMM	0.10%
(Lavrentyeva et al., 2017)	ASVspoof 2017	Log Power Magnitude + CQT, Log Power Magnitude + FFT	CNN + RNN (Convolutional Neural Network + Recurrent Neural Network)	6.73%
(Shim et al., 2018)	ASVspoof 2017	DNN Extracted Features	DNN (Deep Neural Network)	9.56%
(Yang et al., 2018)	ASVspoof 2015 ASVspoof 2017	eCQCC	DNN	ASVspoof 2015- 0.04% ASVspoof 2017-13.38%
(Cai et al., 2019)	ASVspoof 2019	CQCC, LFCC, IMFCC	DNN, ResNet (Residual Network)	0.66%
(Kumar & Aggarwal, 2020d)	ASVSpoof 2019	CQCC, LFCC, IMFCC, LFBC	Time-Delay Shallow Neural Network	5.7%

3 Methodology

The proposed system, as illustrated in Figure 1, outlines a framework for spoof detection using the ResNet152 model that leverages deep learning techniques. The process begins with data collection from the ASVSpoof 2019 datasets, to ensure an adequate amount of data for model training. Following data collection, the raw audio files undergo a preprocessing stage to address noise and anomalies. This refined audio data is then transformed into spectrogram images, which serve as input for feature extraction. Spectral features derived from the spectrograms are subsequently fed into a ResNet152 classification model. ResNet152, a deep convolutional neural network with 152 layers and ReLU activation functions, effectively classifies the input audio as either Genuine (G) or Spoof (S).

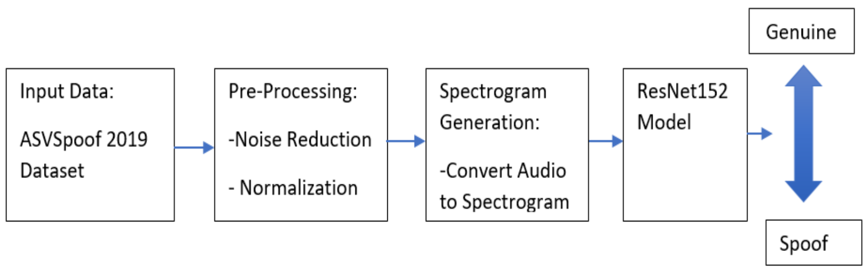


Figure 1: The Framework of Spoof Detection using the ResNet152 Model

Dataset Collection

For a DL model to be run effectively, it needs to have sufficient data for audio spoof detection, the datasets we used are ASVSpoof 2019. Table 2. depicts the overall summary of the datasets being used in this paper (Yamagishi et al., 2019).

ASVSpoof2019: The ASVspoof 2019 database is built on the Voice Cloning Toolkit (VCTK) corpus (Cai et al., 2019). It was constructed by downsampling the 107 speakers' utterances (46 males and 61 females) to 16 kHz at 16 bits/sample. The ASVspoof 2019 database consists of two partitions for assessing logical access (LA) and physical access (PA) scenarios. Both these databases are partitioned into training, development, and evaluation sets which consists of speech from 20 speakers (8 male, 12 female), 10 (4 male, 6 female), and 48 (21 male, 27 female) respectively.

Table 2: Dataset Collection Summary

Dataset	No. of Males	No. of Females	Total Speakers	Dataset Link
ASVspoof 2019	46	61	107	https:/datashare.ed.ac.uk/handle/10283/3336

Pre-processing

The primary objectives of audio signal preprocessing are to improve the quality of the signal, remove unwanted noise, and prepare the data for further analysis or processing. Normalization technique is used for preprocessing, which adjusts the amplitude of the audio signal to a standardized range, often between -1 and 1, to prevent clipping and ensure optimal dynamic range. Following the loading of the audio files, this creates an audio time series as a NumPy array with a 22 KHz mono default Sample Rate (SR). Resampling at 44.1 KHz, to match the requirements of the subsequent processing or analysis steps (Gibert et al., 2016).

Spectrogram

After completing the pre-processing stage, the next step involves converting the raw audio data into spectrogram images for further analysis (convolutional-neural-networks, 2021; vgg19-architecture; introduction - architecture of alexnet, 2024). It allows us to observe variations in energy levels at specific frequencies and how they change over time. Typically depicted as a heatmap, the spectrogram's intensity is represented by varying colours or brightness levels (Kim et al., 2018). Generating a traditional spectrogram is a straightforward process. The pre-processed frames are initially subjected to the Short-time Fourier Transform (STFT). To reduce artifacts, these frames are passed through a 256-point Hanning Window at a 0.5 skip rate. Equation (1) is used to formulate the spectrogram, where the magnitude squared of the STFT coefficients represents the power spectrum.

$$Spectrogram(t, \omega) = |STFT(t, \omega)|^2 \tag{1}$$

To employ the decibel (dB) scale rather than the amplitude scale, we go one step further in this study and apply Equation (2) to convert the traditional spectrogram into a log scale. Figure 2 displays the conversion of the spectrogram.

$$Spectrogram(t, \omega)|_{dB} \;=\; 20log_{10}\left(\frac{|STFT(t,\omega)|}{2x10^{-5}}\right) \tag{2}$$

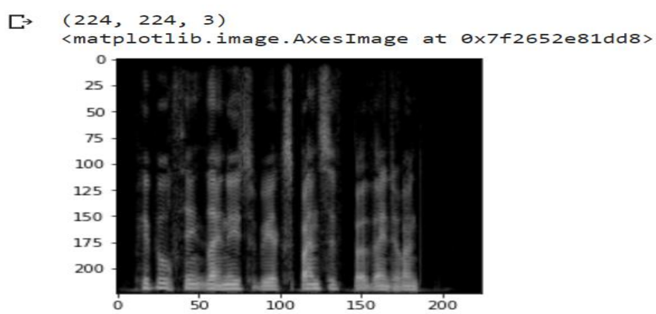


Figure 2: Spectrogram Converted Image

Classification Model

Once all features are extracted from the obtained spectral images, we now pass to the classifier model in which here we use ResNet152 layered neural network. Figure 3. depicts the fundamental structure of ResNet152 (Pustokhin et al., 2023).

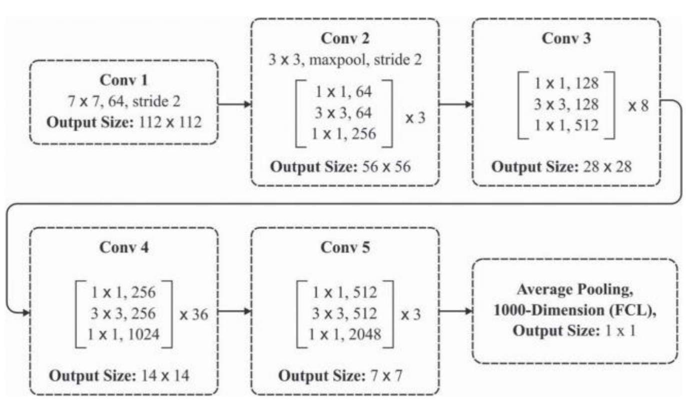


Figure 3: Architecture of ResNet 152 - Layer (Pustokhin et al., 2023)

Spectrograms are now passed over to a 152-layer ResNet network. The core idea of the ResNet architecture is the introduction of residual blocks, which enable the network to learn residual functions with reference to the layer inputs, rather than directly learning the underlying transformations. This approach helps to alleviate the vanishing gradient problem, a common issue in deep neural networks, and allows for the training of considerably deeper models. The residual block consists of a series of convolutional layers, batch normalization, and ReLU activation, with a shortcut connection that bypasses the main transformation. The shortcut connection allows the input of the block to be added to the output of the main transformation, forming a residual connection. As illustrated in Figure 4., Let $H(x)$) is the residual mapping that is used to construct a residual learning block.

About this ResNet block, $H(x) = F(x) + x$ is calculated. "Shortcut connections" in feedforward neural networks enable them to recognize the formulation of $F(x) + x$. Without the need for any further parameters, the shortcut connections use identity mapping to merge the stacked layer's input and output. Gradients can therefore readily flow back, enabling substantially more layers and faster training. More identity links are added to the network in a freshly improved version of ResNet that is being proposed.

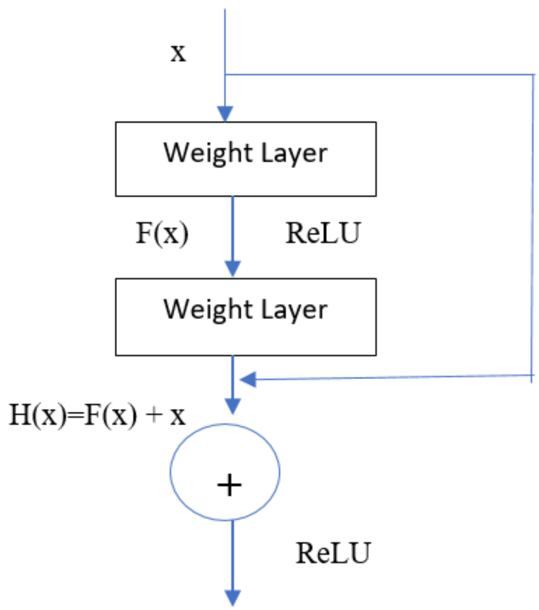


Figure 4: Residual Learning Block

Identity Block: This can be shown as;

$$Y = F(x, W_i) + x \quad (3)$$

Where x and Y are the layers under consideration's input and output vectors. Equation (3), We can refer to the trained residual mapping as the function $F(x, W_i)$. Consequently, the dimensions of x and F in the identity block are equal. The first part is a 2D convolutional layer with a stride of (1,1) and a filter size of (1×1). Batch Normalization performs the channel axis normalization, and the nonlinear activation function is computed using the ReLU function. Similar to the first component, the second one has a different filter size (F x F) (Alzantot et al., 2019).

Convolutional Block: This block has separate input and output dimensions. In equation (4), the dimensions between x and F are resized using linear projection using the shortcut connections:

$$Y = F(x, W_i) + W_{sx} \quad (4)$$

In this shortcut, the input "x" is enlarged to line up with the main route. The 2D convolutional layer has a stride of (s,s) and a filter size of (1×1) depending on the output dimensions. Finally, the modified shortcut and the output of the main path are combined. The main advantage of the revised shortcut is that it can handle the vanishing gradient problem. It guarantees that the upper layer will always function in the same class as the lower layer under all conditions by teaching the model an identity function (cifar 10 dataset, 2021).

4 Performance Evaluation and Implementation

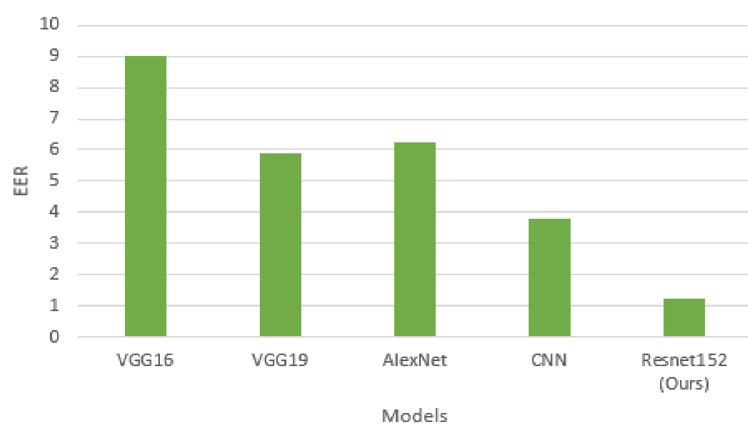
We implement our system in Ryzen 7 5800x processor with 32GB DDR4, NVIDIA GeForce RTX 3080 10GB, SSD of 500GB with Windows 10 OS. Here we evaluated our model under performance measures such as Equal Error Rate, t-DCF and Accuracy. Also, we compare our model with existing models such as VGG16, VGG19, AlexNet, and CNN (datascience, 2019; convolutional-neural-networks, 2021; vgg19-architecture; introduction - architecture of alexnet, 2024).

Tandem detection cost function, or t-DCF, is the new competition primary metric for ASVSpoof 2019 (Kinnunen et al., 2018). It was suggested as a trustworthy grading system to assess the combined effectiveness of CMs and ASV. Equal Error Rate, or EER, is the additional metric that is employed. When the rates of false alarms (false positives) and misses (false negatives) equalize, it is known as the EER. A comparative analysis of the performance measures of various deep learning models with our proposed model on ASVspoof 2019 dataset is shown in Table 3.

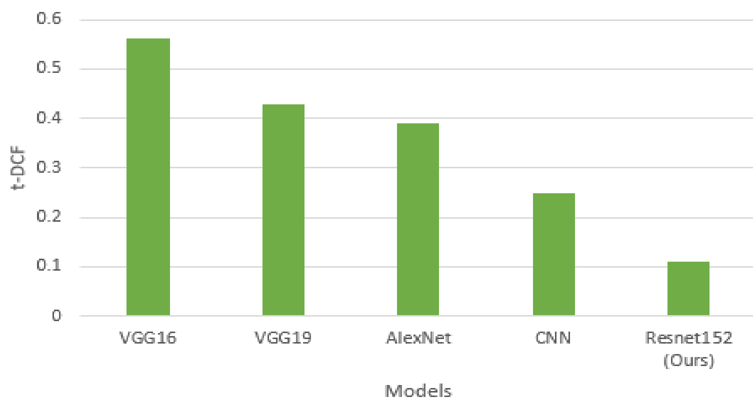
Table 3: Shows the Performance Measures of Various Models with our Proposed Model Under ASVspoof 2019 Datasets

Models	Datasets	EER	t-DCF	Accuracy
VGG16	ASVSpoof2019	8.99	0.56	80%
VGG19		5.90	0.43	81.9%
AlexNet		6.25	0.39	87%
CNN		3.78	0.25	90%
ResNet152(Ours)		1.23	0.11	92.2%

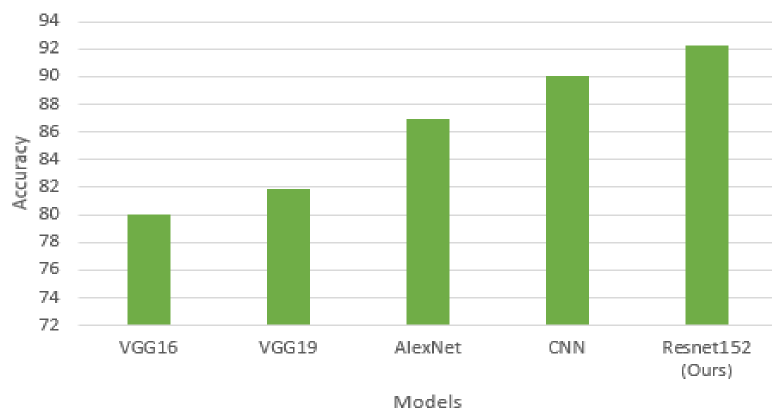
Figure 5. (a, b, c) depict the graphical representation of various models with our proposed model under the ASVSpoof2019 dataset in which our proposed model has the highest accuracy, lowest EER, and t-DCF measure with 92.2%, 1.23, and 0.11.



(a)



(b)



(c)

Figure 5: a) EER vs Models, b) t-DCF vs Models, c) Accuracy vs Models under ASVSpoof2019 Dataset

Figure 6. depict the testing accuracy and also the accuracy when these are applied in real-world instances (inferences) of the ResNet152 model.

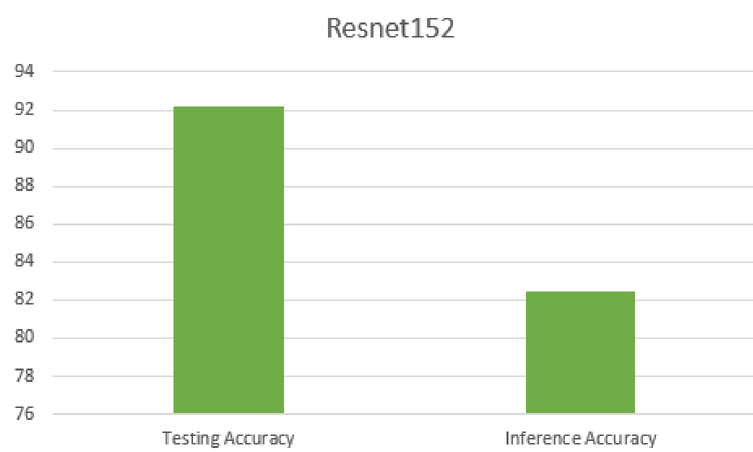


Figure 6: Testing and Inference Accuracy of the ResNet152 Model

5 Conclusion

This paper delves into the impact of deep learning on audio spoof detection, highlighting its potential to accurately discern genuine speakers. Our research leverages popular datasets like ASVSpoof2019, meticulously processed and transformed into spectrogram images. We extract crucial spectral features that empower the model for effective analysis and performance. The cornerstone of our approach is ResNet152, a powerful deep-learning architecture. This model excels at classifying audio as genuine or spoofed with exceptional accuracy. In comparison to other models, ours displays a significantly lower Equal Error Rate (EER), reduced Target DCF (t-DCF), and demonstrably higher accuracy. These metrics underscore the robustness and effectiveness of our proposed system. Furthermore, our experimentation across various datasets consistently reveals the superiority of ResNet152 in audio spoof detection. This finding solidifies its position as a valuable tool for mitigating security risks associated with voice manipulation.

References

[1] Alzantot, M., Wang, Z., & Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. http://dx.doi.org/10.21437/Interspeech.2019-3174

[2] Boulkenafet, Z., Komulainen, J., & Hadid, A. (2015). Face anti-spoofing based on color texture analysis. *In IEEE International Conference on Image Processing (ICIP)*, 2636-2640.

[3] Cai, W., Wu, H., Cai, D., & Li, M. (2019). The DKU replay detection system for the ASVspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion. https://doi.org/10.48550/arXiv.1907.02663

[4] Gibert, K., Sànchez–Marrè, M., & Izquierdo, J. (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. *AI Communications*, *29*(6), 627-663.

[5] Gümüş, A. E., Uyulan, Ç., & Guleken, Z. (2022). Detection of EEG Patterns for Induced Fear Emotion State via EMOTIV EEG Testbench. *Natural and Engineering Sciences*, *7*(2), 148-168.

[6] Hanilçi, C., Kinnunen, T., Sahidullah, M., & Sizov, A. (2015). Classifiers for synthetic speech detection: A comparison. *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, 2057-2061.

[7] https://iq.opengenus.org/vgg19-architecture/

[8] https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c

[9] https://www.analyticsvidhya.com/blog/2021/03/introduction-to-the-architecture-of-alexnet/

[10] https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/

[11] https://www.analyticsvidhya.com/blog/2021/06/understanding-ResNet-and-analyzing-various-models-on-the-cifar-10-dataset/

[12] Kim, D., Sung, T. T., Cho, S. Y., Lee, G., & Sohn, C. B. (2018). A single predominant instrument recognition of polyphonic music using CNN-based timbre analysis. *International Journal of Engineering & Technology*, *7*(3.34), 590-593.

[13] Kinnunen, T., Evans, N., Yamagishi, J., Lee, K. A., Sahidullah, M., Todisco, M., & Delgado, H. (2017). Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training*, *10*(1508), 1508.

[14] Kinnunen, T., Lee, K. A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., & Reynolds, D. A. (2018). t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. https://doi.org/10.48550/arXiv.1804.09618

[15] Kinnunen, T., Wu, Z. Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4401-4404.

[16] Kumar, A., & Aggarwal, R. K. (2020d). A time delay neural network acoustic modeling for hindi speech recognition. *In Advances in data and information sciences*, 425–432.

[17] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. *In Interspeech*, 82-86.

[18] Pal, M., Paul, D., & Saha, G. (2018). Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language*, *48*, 31-50.

[19] Pustokhin, D. A., Pustokhina, I. V., Dinh, P. N., Phan, S. V., Nguyen, G. N., Joshi, G. P., & K, S. (2023). An effective deep residual network based class attention layer with bidirectional LSTM for diagnosis and classification of COVID-19. *Journal of Applied Statistics*, *50*(3), 477-494.

[20] Scardapane, S., Stoffl, L., Röhrbein, F., & Uncini, A. (2017). On the use of deep recurrent neural networks for detecting audio spoofing attacks. *In IEEE International Joint Conference on Neural Networks (IJCNN)*, 3483-3490.

[21] Shim, H. J., Jung, J. W., Heo, H. S., Yoon, S. H., & Yu, H. J. (2018). Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. *In IEEE Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, 172-176.

[22] Srinivasa Rao, M., Praveen Kumar, S., & Srinivasa Rao, K. (2023). Classification of Medical Plants Based on Hybridization of Machine Learning Algorithms. *Indian Journal of Information Sources and Services*, *13*(2), 14–21.

[23] Udayakumar, R., Anuradha, M., Gajmal, Y. M., & Elankavi, R. (2023). Anomaly detection for internet of things security attacks based on recent optimal federated deep learning model. *Journal of Internet Services and Information Security*, *13*(3), 104-121.

[24] Verkholyak, O., Dvoynikova, A., & Karpov, A. (2021). A Bimodal Approach for Speech Emotion Recognition using Audio and Text. *Journal of Internet Services and Information Security*, *11*(1), 80-96.

[25] Wong, S. K., & Yiu, S. M. (2020). Location spoofing attack detection with pre-installed sensors in mobile devices. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *11*(4), 16-30.

[26] Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., & Nautsch, A. (2019). Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. https://doi.org/10.7488/ds/2555, Online source: https://datashare.ed.ac.uk/handle/10283/3336

[27] Yang, J., Das, R. K., & Li, H. (2018). Extended constant-Q cepstral coefficients for detection of spoofing attacks. *In IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1024-1029.

[28] Zhao, Y., Togneri, R., & Sreeram, V. (2018). Spoofing Detection Using Adaptive Weighting Framework and Clustering Analysis. *In Interspeech*, 626-630.

Author Biography

M. Selin, Research Scholar, Department of Computer Applications, Cochin University of Science and Technology, Kerala, India. She has 20 years of teaching experience. Her area of interest includes speaker verification and antispoofing.

Dr.K. Preetha Mathew, Professor, Department of Computer Science and Engineering, Cochin University College of Engineering Kuttanad, Kerala, India. With 24 years of teaching and 4 years of research experience, her areas of expertise include cryptography and network security. She secured the first rank in her M.Tech in Computer Science and Information Sciences. Dr. Preetha Mathew has a significant number of conference and journal publications to her credit.