ISBN 979-8-3503-1512-7

Audio Deepfake Detection using Data Augmented Graph Frequency Cepstral Coefficients

[1]Mohit Dua, [2]Swati Meena, [3]Neelam, [4]Amisha, [5]Nidhi Chakravarty
[1,2,3,4,5]Department of Computer Engineering, National Institute of Technology, Kurukshetra
[1]er.mohitdua@nitkkr.ac.in, [2]swatimeena640@gmail.com, [3]kambojneelam696@gmail.com, [4]amisha50jangir@gmail.com, [5]nidhichakravarty03@gmail.com

2023 International Conference on System, Computation, Automation and Networking (ICSCAN) | 979-8-3503-1512-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCAN58655.2023.10395679

***Abstract* — Automatic speaker verification (ASV) systems serve an important role in identifying speakers in a variety of domains by enabling authentication, convenience, fraud detection, personalization, and forensic applications. The demand for ASV systems originates from how simple and effective speech biometrics may be. The growing popularity of such applications raises concerns about the growing possibility of speech attack. The purpose of this research is to identify audio spoofing attacks in an ASV system. The suggested model has a front-end and a back-end. The front-end has two features: Gammatone Cepstral Coefficients (GTCC) and Graph Frequency Cepstral Coefficients (GFCC) based on Spectrograms. Four machine learning models are utilised in the backend: Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and K-nearest Neighbour (KNN), as well as one deep learning model named Long Short-Term Memory (LSTM) and a transfer learning based pertained ResNet 50 model. The Logical Access (LA) partition of ASVspoof2019 is used for training, whereas the Deepfakes (DF) portion of ASVspoof 2021 is used for testing. To address the issue of dataset imbalance, methods such as SpecAugment and Speed perturbation are used to extracted features, particularly GTCC features. For deep fake detection, the suggested model, which combines GFCC with pretrained ResNet 50, obtains an outstanding Equal Error Rate (EER) of 1.78% and a tandem-Detection Cost Function (t-DCF) of 0.0458 min.**

Keywords— ***ASV; GTCC; EER; LA; GSP, ResNet50, LSTM***

I. INTRODUCTION

Consider a scenario where a friend asks you for money urgently, and you promptly send it to them. However, later you realize that their phone was stolen, and the call was made by an imposter. Did you lose your money? But there's hope! If you have an ASV system on your phone and it detects that the voice on the call does not match your friend's voice, then you can be saved. This highlights the potential of ASV as a technology that could protect against fraudulent voice-based transactions and impersonation, making it a compelling subject for further research. However, due to the increasing number of attacks on ASV systems, people may be hesitant to use them. The main motivation of this research paper is to focus on the detection of such attacks and develop methods to enhance the security and reliability of ASV systems. Majorly focusing on the LA attacks and DF attacks.

An ASV system evaluates if the input speech signal was generated by the authentic user or by the imposter in order to get access to the real user's account.

Figure 1 depicts the ASV system's architecture. The ASV system in this proposed work is separated into two parts: the front end and the rear end. The front end is only concerned with extracting features from audio samples, making audio processing much easier. In contrast, the backend mainly depends on decision-making algorithms to evaluate whether the audio is authentic or faked. The user is authenticated based on the decision of the backend model. It is extensively used for person authentication, and tries to validate the stated identity of a speaker [1]. Even though the technology has advanced significantly in recent years, however, tests have shown that these systems still are vulnerable to spoofing, commonly known as a spoof attack [2], miming [3], text-to-speech (TTS), replay [4], and voice-conversion (VC) spoofing attacks[5][6].

Deep learning has shown promising results in various applications, including audio speaker verification. Deep learning models can be used to extract features from the recorded audio signals and perform speaker verification tasks. It's worth noting that deep learning models require a large amount of labeled data to train effectively. The accuracy of speaker verification is also affected by the quality of the recordings. Overall, deep learning has demonstrated significant promise for audio speaker verification.

A. Related Work

Feature extraction from audio signals is a key aspect of the front end of an ASV system. Sharma et al. [7] emphasize the importance of feature extraction in their work. Tiwari et al. also utilize the Mel Frequency Cepstrum Coefficient (MFCC) feature for developing a text-dependent speaker identification system. In a different study, Fathima et al. [8] explored the use of GTCCs, which inspired by biological mechanisms. A comparison between MFCC features and GTCC features for speaker identification reveals that GTCC yields significantly higher classification accuracies compared to MFCC. These derived attributes make it easier to investigate the differences between audio. Kwak et al. [9] proposed GFCC (graph frequency cepstral coefficient) features based on GSP with a light convolutional neural network for identifying replay attacks. They proved it to be one of the strongest front-ends when compared to other existing systems on the ASVspoof 2019 physical access corpus. Graph Signal Processing (GSP) as Liu et al. [10] proposed shows a better correlation between audio samples and explored hidden information from speech. The graph is a representation that describes the geometric structure of a data field

ISBN 979-8-3503-1512-7

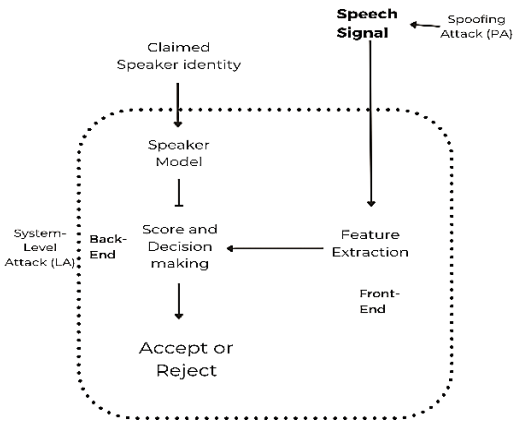

Figure 1: Architecture of Automatic Speaker Verification system

It contains high-dimensional data (signals) on its vertices and edges , with weights on the edges representing the similarity between connected vertices. Meng et al. [11] worked on 3-D Mel spectrograms using CNNs to classify the emotion from the audio provided. This was one of the first of its kind. Nucci et al. [17] used fuzzy clustering to build the tree for population partitioning , which means that a speaker might belong to many groups at each level. Wan et al. [18] demonstrated a text-independent speaker verification system based on SVMs with score-space kernels. Dua et al. [13] used LSTM for classification on the backend with the GTCC features [12][24]. Deep Learning has been more prevalent lately. Sadhu et al. [14] used Convolutional Neural Network (CNN), LSTM , and a hybrid of both to find the distinction between bonafide and spoof audio samples. In [15], Convolutional Neural Networks

have been the most prominently used algorithm in the ASV systems as well as the most efficient DL model. In [16], authors proposed a noise robust model using hybrid MFCC, GTCC features and Time Delay Neural Network for spoof detection. The suggested work in this paper was inspired by the above approaches. The models used at the backend were LSTM , ResNet , and a transfer learning approach combining the Residual Neural Networks. The frontend is mainly composed of feature extraction methods such as GTCC , and spectrograms generated by GSP. To detect LA and DF attacks, the ASV spoof corpus has been used. As a result, the following are the paper's contributions:

1. In this work , we have used GTCC features and GFCC spectrogram at the front end. To solve the problem of data imbalance , SpecAugment and Speed perturbation data augmentation technique have been applied to GTCC features.
2. Four machine learning algorithms such as KNN, SVM, DT and LR, one deep learning model ; LSTM and one Transfer learning based pretrained ResNet 50 have been used at backend.
3. To evaluate the performance of the suggested model EER and t-DCF evaluation metrics has been used.
4. The training audios in this study were chosen from the ASVspoof 2019 LA division, while the audios used to test the model against deepfake attacks were chosen from the ASVspoof2021 dataset.

II. Proposed Method

The proposed architecture's flow is depicted in Figure 2. The ASV system is divided into three phases: Front end, Backend, and Decision. In the Front-end phase, feature extraction is performed using GTCC and GFCC features. Initially, GTCC features are extracted, and data augmentation techniques like SpecAugment and Speech Perturbation are applied to enhance these features. Additionally, the GTCC feature is passed through the GSP unit to extract the GFCC spectrogram. Moving to the Backend phase, the GTCC features are utilized by several models, including DT, KNN, SVM, LR, and LSTM. On the other hand, the GFCC spectrogram is applied to pretrained ResNet50. At the Decision level, decision is done on the basis of a threshold value whether the audio is real or spoofed.

A. Feature Extraction

This section provided the overview of the proposed feature extraction techniques.

GTCC: The GTCC signal is divided into small frames, which provides a static view of a dynamic signal and enhances the performance of the feature extraction process. In GTCC, we utilize the properties of the Gammatone Filter, which are similar to the human response to various frequencies. The filter helps to remove irrelevant frequencies from the signal. Initially, the GTCC output consisted of 43 coefficients for a single audio, denoted as N X 43, where N is the number of audios. This output was subsequently reduced to N X 23 by applying Graph Signal Processing (GSP). Only the positive frequencies and the DC component of a signal are included in the computation of the Fast Fourier Transform (FFT). This is due to the fact that the FFT result is symmetrical around index 43//2. The generated spectrogram S has the shape (43//2+1, n), where n is the number of frames in the GTCC data. This step is required to see and analyze the frequency content of the data across time. Figure 3 shows extraction of the features. The steps to build GTCC for a signal are as follows:

- Audio Input: The system takes in an audio signal.
- Pre-emphasis: A high-pass filter is used to amplify the high-frequency components of the signal and attenuate low-frequency noise.
- Gammatone Filter bank: The signal is filtered through a bank of Gammatone filters that emulate the frequency selectivity of the human auditory system.
- Log Compression: Each output from the gammatone filters is compressed logarithmically to simulate the nonlinear response of the human auditory system.
- Discrete Cosine Transform (DCT): The log-compressed filter outputs undergo a DCT to generate a set of cepstral coefficients.
- GTCC Output: A subset of the DCT coefficients is chosen to create the GTCC feature vector. These are generally the lower-order coefficients that capture the signal's overall spectral envelope.
- This GTCC output is then used in Graph signal processing to extract graph frequency cepstral coefficient.

ISBN 979-8-3503-1512-7

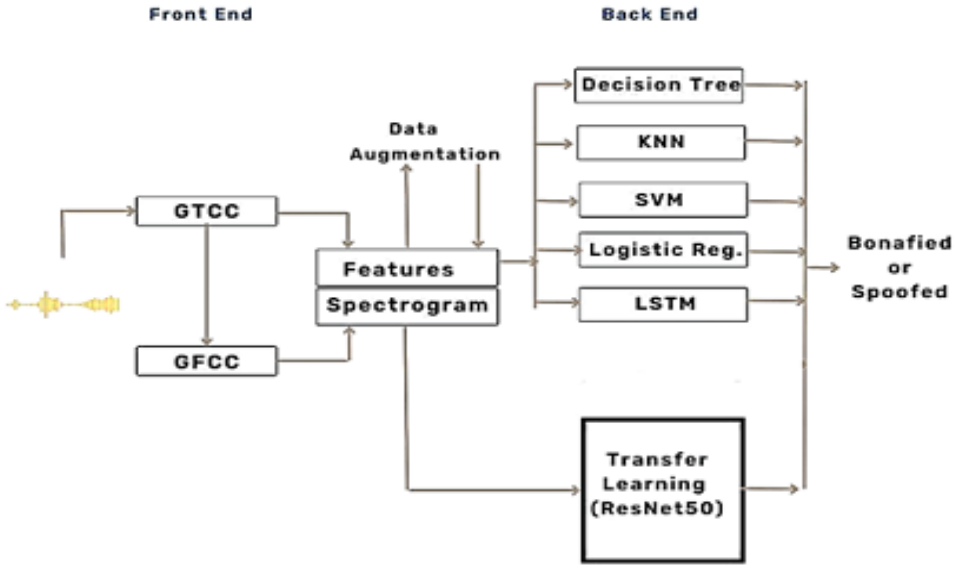

Figure 2: Flow of Proposed Methodology

Figure 3: Steps to extract GTCC features

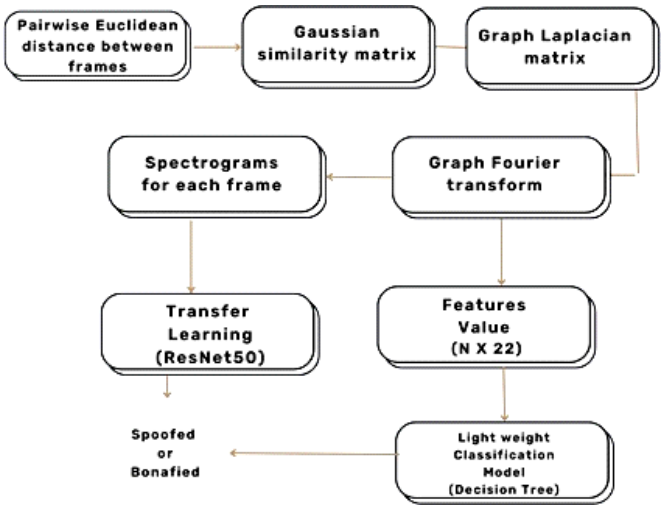

Figure 4: Steps to extract GFCC spectrograms

Graph Signal Processing (GSP): GSP is a field of study concerned with the analysis, processing, and manipulation of signals defined on graphs. In this context, signals are represented as functions on graph nodes and can capture a variety of phenomena such as social interactions, traffic flows, or brain activity. Its goal is to create tools and methods for analyzing and processing these signals in order to better understand the underlying systems they represent. The graph Fourier transform is used in graph signal processing to analyze the structure and properties of signals defined on graphs. This transform breaks signals down into frequency components [19]. We have proposed GFCC features. Figure 4 shows the steps to extract GFCC features.

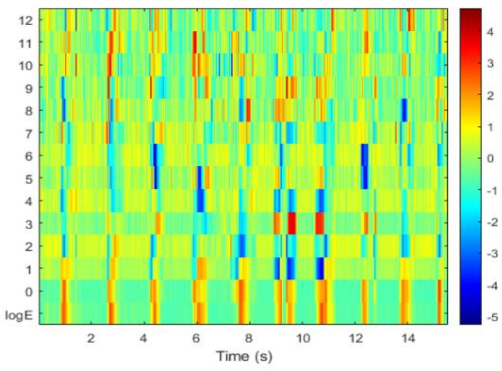

Figure 5: GFCC Spectrogram Representation

B. Data Augmentaion

We performed data augmentation on the retrieved GTCC coefficients to boost the diversity of the training data. This method is more computationally efficient than enhancing the raw audio signals. To produce fresh data from the original GTCC coefficients, we applied three alternative augmentation techniques: SpecAugment, and Speed perturbation. This procedure entailed randomly perturbing several features of the GTCC data, such as the time and frequency dimensions, in order to generate variations of the original data. By doing so, we hoped to avoid overfitting and increase our models' generalization performance throughout training.

- **Speed perturbation:** This is accomplished by altering the speed or tempo of an audio signal. The basic idea behind speed perturbation is to change the playback rate of an audio signal while keeping its pitch constant. We can create new training examples that capture variations in speech rate and tempo by varying the speed of the audio signal, which can help improve the robustness of machine learning models to such variations.

- **SpecAugment:** SpecAugment involves masking and warping the spectrogram representation of an audio signal.

C. Backend Classification

This section gives the overview of the techniques applied to implement backend of the suggested model.

- **Transfer Learning:** Technique that uses a pre-trained model to solve a new problem. Fine-tuning the pre-trained model involves replacing its final layers and re-training them on the new data.

This allows us to adapt the pre- trained model to the new task and improve its accuracy with less training data . The ability to transfer knowledge from one task to another has the potential to revolutionize the way we approach machine learning and make it more accessible to a wider audience.

- **Decision Trees (DT):** Tree-Structured classifiers that are used to solve classification problems. In the decision trees, leaf nodes are outcomes, branches are decision rules, and internal nodes are features of a dataset.

- **Support Vector Machine (SVM):** Generates a decision boundary called a hyperplane to classify new data points into different classes. SVM uses extreme points to create a support vector.

- **Logistic Regression (LR):** Model used for predicting the output of categorical dependent variables in the form of probabilistic values between 0 and 1 using sigmoid functions.

- **K-Nearest Neighbours (KNN):** A non-parametric and lazy learner algorithm that categorizes data into different classes based on the similarity between new data and existing data.

- **LSTM Architecture:** LSTM, is a type of RNN that addresses the problem of vanishing gradients, which causes data shared at the beginning of the input sequence to disappear and results in poor prediction. LSTM maintains both long-term memory (LTM) and short-term memory (STM) in the memory cell to improve performance.

III. EXPERIMENTAL SETUP AND RESULTS

This section gives the details about the dataset used and experiment performed during the work.

A. Dataset

This work uses the ASVSpoof2019 Logical access (LA) and ASVspoof2021 Deepfake (DF) dataset. It is a collection of genuine and spoofed speech signals that were recorded under various conditions. The ASVspoof 2019 database consists of two sections, one for logical access (LA) and the other for physical access (PA) scenarios. The objective of the new DF ASVspoof2021 is to evaluate the effectiveness of anti-spoofing detection solutions in identifying manipulated speech data that has been compressed in different ways and posted on various online platforms. Table 1 shows details of dataset used.

TABLE I : DATASET USED

Dataset ASVspoof	Attack Type	Train	Development	Evaluation
2019	LA	25380	24844	-
2021	DF	-	-	20,000

B. Experments and Results

Table 2 shows the result of the proposed model. Two experiments have been done. In first experiment GTCC features applied to DT, KNN, SVM, LR and LSTM and produced 2.79, 2.97, 2,3.18, and 1.96 % EER, respectively. In second experiment GFCC spectrograms have been applied to pretrained ResNet50 model and produced 1.78% EER. From result we can say that GFCC spectrogram outperformed the GTCC feature.

ISBN 979-8-3503-1512-7

TABLE II : RESULT OF THE PROPOSED MODEL FOR DEEPFAKE ATTACK

Feature Set	Backend Model	EER	Acc[1]	Prec[2]	Recall	F1[3]	Min t-DCF
		(%)					
GTCC	DT	2.79	98.3	97	98	98.2	0.085
	KNN	2.97	98	97	97	98.4	0.087
	SVM	2	98	98	98	98.1	0.059
	LR	3.18	96	96.2	96	96	0.102
	LSTM	1.96	98	98	98	98.6	0.0658
GFCC Spectrogram	ResNet50	1.78	98	99	99	99.1	0.0458

IV. Comparative analysis with existing Techniques

We compared our system against existing models for detection of audio deepfake attack to evaluate the performance of our suggested method. Table 3 shows the comparative results. We employed the ASVspoof 2019 LA portion's training and ASVspoof2021 DF set for testing the proposed model and all the comparative method of spoof attack.

TABLE III: COMPARATIVE ANALYSIS WITH EXISTING TECHNIQUES

Author/ Year	Front end Technique	Backend Model	Dataset used	Evaluation Parameter					
				Acc[1]	Prec[2]	Recall	F1[3]	EER	t-DCF
Gharde et al, [20], 2022	Mel spectrogram	ResNet50V2	ASVspoof 2019	97.1	×	×	×	×	×
Fathan et al, [21], 2022	Wavelet CNN, VGG16	ASVspoof2019, 2021	ASVspoof2019,2021	×	×	×	×	0.39	×
Xue *et al* [22],2022	Sub band fusion of F0, imaginary and real spectrograms	squeeze-and-excitation ResNet	ASVspoof 2019	×	×	×	×	1.21	0.0358
Doan et al, [23],2023	LFCC	RawNet2,	ASVspoof 2019 and 2021	×	×	×	×	8.75	38.93
Proposed	GFCC spectrogram	ResNet50	ASVspoof2019,2021	98	99	99	99.1	1.78	0.0458

Note: Accuracy[1], precision[2], F1-score[3]

V. Conclusion

This paper proposes a GTCC and GFCC spectrogram-based feature which takes GTCC features as input. In first experiment, augmented GTCC features applied to four ML models; DT, KNN, SVM, LR. Out of all these four model SVM produced the min EER that is 2 %. Along with these ML model LSTM is also applied and produced 1.96 % EER. From this we can say that GTCC produced better EER with LSTM. In the second experiment, GTCC features applied to GSP unit to extract GFCC spectrogram. These spectrograms have been applied to ResNet50 model and achieved 1.78 % EER. The results also show that using spectrogram with pretrained Deep Convolutional Neural Network model improve the speaker verification other assessment metrics. Future research should focus on developing more efficient and robust methods for detecting and preventing these attacks.

References

[1] D. A. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian Mixture speaker models,” Speech Audio Process. IEEE Trans., vol. 3, pp. 72–83, Feb. 1995, doi: 10.1109/89.365379.

[2] R. Tolosana, M. Gomez-Barrero, C. Busch, and J. Ortega-Garcia, “Biometric presentation attack detection: Beyond the visible spectrum,” IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 1261–1275, 2019.

[3] Y. W. Lau, M. Wagner, and D. Tran, “Vulnerability of speaker verification to voice mimicking,” in Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., 2004, pp. 145–148.

[4] Z. Wu, S. Gao, E. S. Cling, and H. Li, “A study on replay attack and anti-spoofing for text-dependent speaker verification,” in Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014

Asia-Pacific, 2014, pp. 1–5, doi: 10.1109/APSIPA.2014.7041636.

[5] X. Wang et al., “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” arXiv Prepr. arXiv1904.05441, 2019.

[6] N. Chakravarty and M. Dua, “Spoof Detection using Sequentially Integrated Image and Audio Features,” Int. J. Comput. Digit. Syst., vol. 13, no. 1, p. 1, 2023.

[7] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020.

[8] Fathima, R., & Raseena, P. (2013). Gammatone cepstral coefficient for speaker Identification. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(1), 540-545.

[9] Kwak, I. Y., Kwag, S., Lee, J., Huh, J. H., Lee, C. H., Jeon, Y., ... & Yoon, J. W. (2021, January). Resmax: Detecting voice spoofing attacks with residual network and max feature map. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4837-4844). IEEE

[10] Cai, W., Cai, D., Liu, W., Li, G., & Li, M. (2017, August). Countermeasures for Automatic Speaker Verification Replay Spoofing Attack: On Data Augmentation, Feature Representation, Classification and Fusion. In Interspeech (pp. 17-21).

[11] Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE access, 7, 125868-125881.

[12] Joshi, S., & Dua, M. (2022, May). LSTM-GTCC based Approach for Audio Spoof Detection. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (Vol. 1, pp. 656-661). IEEE.

[13] Mittal, A., & Dua, M. (2021). Automatic speaker verification system using three dimensional static and contextual variation-based features with two-dimensional convolutional neural network. International Journal of Swarm Intelligence, 6(2), 143-153.

[14] Dua, M., Sadhu, A., Jindal, A., & Mehta, R. (2022). A hybrid noise robust model for multi-replay attack detection in Automatic speaker verification systems. Biomedical Signal Processing and Control, 74, 103517.

[15] Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 22.

[16] Chakravarty, N., & Dua, M. (2022). Noise Robust ASV Spoof Detection Using Integrated Features and Time Delay Neural Network. SN Computer Science, 4(2), 127.

[17] Hu, Y., Wu, D., & Nucci, A. (2012). Fuzzy-clustering-based decision tree approach for large population speaker identification. IEEE transactions on audio, speech, and language processing, 21(4), 762-774.

[18] Wan, V., & Renals, S. (2005). Speaker verification using sequence discriminant support vector machines. IEEE transactions on speech and audio processing, 13(2), 203-210.

[19] Xu, L., Tian, M., Guo, X., Shan, Z., Jia, J., Peng, Y., ... & Das, R. K. (2022). A Novel Feature Based on Graph Signal Processing for Detection of Physical Access Attacks}}. In Proc. The Speaker and Language Recognition Workshop (Odyssey 2022) (pp. 107-111)

[20]D. Gharde, N. Suryanarayan, and K. S. Srinivas, “Detection of Morphed Face, Body, Audio signals using Deep Neural Networks,” in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022, pp. 1–6.

[21] A. Fathan, J. Alam, and W. H. Kang, “Mel-Spectrogram Image-Based End-to-End Audio Deepfake Detection Under Channel-Mismatched Conditions,” in 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1–6.

[22] J. Xue et al., “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, 2022, pp. 19–26.

[23] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, “BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[24] Chakravarty, N., & Dua, M. (2023). Data augmentation and hybrid feature amalgamation to detect audio deep fake attacks. Physica Scripta, 98(9), 096001.