Volume: 16 Issue: 05 Pages: 6551-6560 (2025) ISSN: 0975-0290

A Deep Learning Fusion Model Leveraging Spectral Features for Audio Deepfake Detection

Nirmal Kaur
Dept. of CSE, UIET, Panjab University, Chandigarh, India
Email : nirmaljul19@gmail.com

Abhishek Dixit
Dept. of CSE, UIET, Panjab University, Chandigarh, India
Email : abhishekdixit4500@gmail.com

Staffy Kingra
Dept. of CSE, SGT University, Gurugram (Haryana), India
Email : staffy_feat@sgtuniversity.org

---ABSTRACT---

Audio deepfakes, a subset of deepfake technology, employ machine learning or deep learning to create deceptive audio content by synthesizing authentic recordings. Such deepfakes not only fosters the dissemination of misinformation but also empowers identity theft while compromising individual privacy. Discerning between counterfeit and authentic audio content poses escalating challenges for digital forensic analysts. The proposed paper develops a robust deep learning model that harnesses fusion approach with a spectrum of diverse audio spectral features to effectively detect deepfake audios. By employing a fusion strategy, the developed model ensembles predictions from two pre-trained networks, CIFAR-10 and ResNet50. Additionally, it capitalizes on a diverse array of spectral audio features- Mel Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), Mel-Spectrogram, and Spectral Centroid, for extraction of crucial details from raw audio data. Accuracy of proposed model is assessed on more recent and widely used FoR dataset having three sub-datasets of 195,000 audio samples. Experimental results reveal that proposed model achieves superior performance, boasting an accuracy of 99.12%, precision of 97.54%, recall of 98.44%, and F1 score of 98.12% when utilizing MFCC feature as compared to other audio features. Moreover, the model undergoes an accuracy-centric quantitative assessment, surpassing eight state-of-the-art audio detection models, including DNN, DeepSonar, STN, TCN, SVM, CNN, KNN, and RF.

Keywords **- Audio Datasets, Audio Deepfake, Deepfake Detection, Machine Learning**

--

Date of Submission: December 21, 2024 Date of Acceptance: January 29, 2025

--

1. INTRODUCTION

Technological advancements in Artificial Intelligence made it easy to produce incredibly lifelike fake audio recordings that can propagate misinformation, sway public opinion, and commit cybercrimes. Several generation techniques have emerged capable of replicating any person's voice with striking realism, presenting a notable challenge to the authenticity of audio-based interactions. This underscores the urgent need for reliable methods to detect deepfake audios. These recordings undergo manipulation through deep learning techniques, called audio deepfakes. Recently, text-to-speech (TTS) [1] and voice conversion (VC) [2] techniques have simplified the creation of synthetic speech. Besides video deepfakes [3][4][5], there has been a notable rise in audio deepfakes competent enough to commit forgeries, enabling individuals to portrait identities of others. Audio deepfakes threat became notably apparent in 2019 when a UK company's CEO fell victim to a scam phone call, leading to significant financial loss of 220,000€. The fraudulent call leveraged audio deepfake technique to convincingly impersonate voice of the CEO of parent company. The intricacies and complexities in distinguishing deepfake audio from authentic recordings present significant detection barriers. For instance, deepfake audio may exhibit subtle intrinsic artifacts like slight fluctuations in pitch, tone, or rhythm, complicating analysis using conventional audio detection techniques. Moreover, when deepfake audio is embedded within longer recordings, detection becomes even more challenging. With recent advancements in GAN (Generative Adversarial Networks) models, forgers are adopting an anti-forensic approach to produce enhanced audio deepfakes, leading to an ongoing arms race between the creation and detection of falsified content, seemingly without end.

Fig. 1 provides a glimpse of deepfaked images, each synthesized through various generation mechanisms. For example, first two images in first row are entirely computer-generated, whereas the third image is produced using StyleGAN [10]. However, remaining images are sourced from deepfaked videos manipulated using Face2Face [11] and NeuralTextures [12] techniques. Sufficient research and development in this area is crucial for preserving the integrity of both audio and video content [6][7][8][9]. Speech synthesis entails the creation of human-like speech through software or hardware programs, serving diverse purposes across domains like text-to-speech (TTS) applications and functioning as personal digital assistants. In the realm of speech synthesis, TTS systems analyze text and produce speech that aligns with the linguistic features of the input text. One advantage of speech synthesis is its ability to offer a variety of accents and voices without the need for pre-recorded human speech samples. For instance, Lyrebird, a notable voice synthesis company, utilizes deep learning models to synthesize up to 1,000 phrases per second. TTS systems greatly depend on the quality of speech used to construct their corpus. However, creating such corpora can be expensive.

Fig. 1: Deepfaked images created through diverse deepfake techniques

Initial research on audio deepfake detection focused on handcrafted feature-based models, requiring manual feature extraction along with significant time overheads. Subsequently, researchers transitioned towards machine learning models. For instance, Rodriguez-Ortega et al. [13] employed logistic regression on the H-Voice dataset [14] to discern fake audio, achieving a detection performance of 98%. With advancements, deep learning models became prominent for their capability to autonomously extract features. Liu et al. [15] conducted an analysis between support vector machines (SVM) and convolution neural networks (CNN), evaluating their effectiveness on a Chinese dataset for detecting fake stereo audios.

However, these techniques mainly revolved around traditional machine learning or basic CNN models, while sidelined the fusion or ensemble approach, which could amalgamate predictions from multiple models to attain greater accuracy. In contrast, the proposed paper implements a late fusion approach, which integrates predictions from two independent deep learning based models. Each of these individual models can operate autonomously, leveraging different spectral audio features. Results reveal that fusion approach optimally exploits the unique characteristics captured by different features and networks, thereby enhancing the overall detection accuracy. Additionally, we compare the accuracy of eight state-of-the-art audio detection classifiers- DNN, DeepSonar, STN, TCN, SVM, CNN, KNN, and RF against the proposed developed model.

Major contributions of the proposed paper are as:

- Deep learning fusion model is developed that ensembles the predictions from two pre-trained networks- CIFAR-10 and ResNet50 to classify real and fake audios.
- Various spectral features- MFCC, CQCC, Mel-spectrogram, Spectral centroid are utilized to extract intrinsic audio features and patterns.
- Developed model is trained and tested on popular Fake or Real (FoR) audio dataset having 195,000 audio samples. Experiments are conducted on three subparts of this dataset.
- Potency of proposed model is analysed on each individual spectral feature, to assess optimal audio feature for deepfake detection.
- Proposed model achieves accuracy (99.12%), precision (97.54%), recall (98.44%) and F1 score (98.12%) when utilizing MFCC feature as compared to other audio features.
- Quantitative accuracy-based evaluation of proposed model outperforms eight state-of-the-art deepfake audio detection models, namely, DNN, DeepSonar, STN, TCN, SVM, CNN, KNN, and RF.

The remainder of the paper is organized as follows. Section 1 provides introduction of audio deepfake and its effects on society. Section 2 illustrates related work on audio deepfake generation and detection methods, along with motivation to contribute to this research area. Afterwards, Section 3 elaborates the proposed methodology, highlighting spectral audio features, audio dataset, and phases of proposed model. Section 4 presents the experimental results obtained, highlighting the accuracy achieved by proposed method against state-of-art techniques. Finally, Section 5 focuses on discussions and conclusions with potential future directions.

2. RELATED WORK

Audio deepfakes synthetically produce audio, frequently generated through machine learning (ML) or deep learning (DL) algorithms, closely mimicking authentic audio recordings. Given their association with numerous illicit activities in recent years, detecting audio deepfakes holds significant importance. By comprehending the techniques employed in their creation, effective detection methods can be devised to mitigate the potential risks linked to their misuse. Conventional acoustic features form the bedrock for characterizing the attributes of audio clips. Derived directly from the raw audio signal, these features encapsulate diverse aspects of the sound waveform, offering quantitative insights into the temporal and spectral content of audio data. In contrast, deep learning features are automatically extracted through deep neural networks on extensive datasets. Table 1 offers a comprehensive summary of different techniques employed for detecting audio deepfakes, delineating their limitations and challenges.

2.1 Machine learning techniques for audio deepfake detection

Machine learning classifiers stand at the forefront of safeguarding the integrity of audio content. Kumar-Singh and Singh [16] introduced a Quadratic Support Vector Machine (Q-SVM) model to distinguish synthetic audio and natural human voices. Authors conducted binary classification of real and synthetic voices, with comparative analysis against other machine learning methods, including Linear Discriminant, Quadratic Discriminant, Linear SVM, weighted K-Nearest Neighbors (KNN). With an accuracy of 97.56% and a misclassification rate of 2.43%, their results showed that Q-SVM model beat other traditional approaches.

Borrelli et al. [17] introduced an innovative approach by combining Support Vector Machine (SVM) with Random Forest (RF) to identify synthetic voices. They leveraged a newly devised audio feature termed Short-Term Long-Term (STLT) and trained their models on Automatic Speaker Verification (ASV) spoof challenge 2019 dataset [18]. The SVM model surpassed RF performance by 71%, highlighting its effectiveness in accurately predicting synthetic voices. Similarly, Liu et al. [15] conducted a comparative analysis between SVM and CNN for detecting synthetic audio amidst genuine recordings. Despite both methods achieving a commendable accuracy of 99%, the study revealed that CNN exhibited superior robustness compared to SVM. This finding underscores the potential of deep learning approaches, particularly CNNs, in effectively discerning synthetic audio from authentic sources. Moreover, recognizing fake audio through traditional machine learning methods can be laborious and susceptible to inconsistencies. Consequently, there has been a notable pivot towards the adoption of DL techniques.

2.2 Deep Learning Techniques for Audio Deepfake Detection

Deep learning techniques with potency to automatically learn intricate patterns and features directly from data, streamline the detection process by eliminating time consuming process of manual feature extraction and extensive pre-processing. M. Ballesteros et al. [19] yielded Deep4SNet, a classification model leveraging 2D CNN architecture to distinguish between imitated and synthetic audio. Deep4SNet achieved an accuracy of 98.5%, signaling promising strides in audio forgery identification. However, its scalability faced constraints due to limitations in handling larger datasets. Subsequent studies, including those by other researchers [20], embarked on comparative analysis of deep learning models for synthetic audio detection. Lataifeh et al. [21] conducted an experimental study that compared performance of CNN and Bidirectional Long Short-Term Memory (BiLSTM) models. They focused on distinguishing real voices from imitators using Arabic Diversified Audio (AR-DAD) dataset [22] comprising Quranic audio clips.

While CNN models exhibited notable accuracy rates, challenges such as overfitting surfaced, indicating the need for further refinement. Apart from traditional methodologies, some researchers [23] turned to scatter-plot images of real and fake audio data to train CNN-based models for binary classification. Despite yielding promising results with an accuracy of 88.9% on Fake or Real (FOR) dataset [24], these approaches highlighted the evolving landscape of deepfake detection. P. RahulT et al. [25] proposed a novel framework targeting fake English-speaking voices, leveraging transfer learning of ResNet-34 model to address the issue of vanishing gradient problem. Despite achieving an impressive Equal Error Rate (EER) of 5.32%, the computational overhead associated with training the deep architecture of ResNet-34 remained a challenge. Similarly, investigations by Khochare et al. [26] explored both feature-based and image-based approaches for classifying synthetically generated faked audio, shedding light on the multifaceted nature of deepfake detection domain. While these endeavours represent significant advancements in the field, challenges still persist. Scalability issues, manual processing requirements, as well as limitations in handling transformed inputs emphasize the necessity of ongoing innovation. Nevertheless, comprehensive surveys [27][28] provide invaluable insights into the evolving landscape of deepfake generation and detection techniques, guiding researchers towards more effective solutions in the ongoing battle against audio manipulations.

3. METHODOLOGY

3.1 Proposed Audio Deepfake Detection Model

Proposed model comprises of three main phases, pre-processing, audio feature extraction, and classification based on fusion, as shown in Fig. 2. In the initial phase, pre-processing with normalization is applied to the input audio waveform. This step scales the audio to a consistent range, ensuring that the extracted features have similar magnitudes. In the second phase, four spectral audio features- MFCC, Mel-spectrogram, CQCC, and Spectral Centroids are used to extract the intrinsic details from the pre-processed audio. These features provide essential information about audio's frequency content, timbre, and other relevant characteristics. Third phase takes extracted audio features and passes them through two independent networks: Convolution CIFAR-10 and ResNet50 network for classification of deepfake and real audio. In this phase late fusion technique is applied using a global pooling layer to aggregate the predictions from these two networks. Fusion strategy allows complementary information extracted from separate models, and thereby enhancing the overall detection performance. On the basis of ensemble information of both networks and learned representations, classifier distinguishes whether audio is real or deepfaked. Detailed description of audio dataset, and various phases of proposed model is explained in the following subsections.

Table 1: Audio deepfake detection techniques

Reference/Year	Audio Deepfake Generation	Technique/Model	Audio Features	Dataset	Limitations
P. RahulT et al. [25]/2020	Synthetic based	ResNet-34	Spectrogram	ASV spoof 2019 [18]	Transformation is required in 2-D before giving input for detection. More training time required.
Kumar-Singh and Singh. [16]/2020	Synthetic based	Q-SVM	MFCC, Mel-Spectrogram	-	Features were extracted manually. Model was not scalable. Extensive labour required for feature extraction.
Borrelli et al. [17]/2021	Synthetic Based	RF, SVM	STLT	ASV Spoof 2019 [18]	Model was not scalable. Needed extensive labour.
Liu.et al. [15]/2021	Synthetic	SVM, CNN	MFCC -	-	Model was not scalable. Error rate is zero indicating CNN model is overfitting.
M. Ballesteros et. al [19]/2021	Synthetic based Imitation based	Deep4SNet	Histogram, Spectrogram, Time domain waveform	H-Voice [14]	Model was not scalable. Data transformation process affects the model.
Khochare et al. [26]/2022	Synthetic based	Feature-based Image-based (CNN, TCN, STN)	Vector of 37 features of audio Mel spectrogram	FoR dataset [24]	Model was not scalable. Needed extensive labour. An image-based methodology was applied. Could not work with inputs converted to STFT and MFCC
S. Camacho et al. [23]/2021	Synthetic based	CNN	Scatter plots	FoR dataset [24]	Model didn't perform well as more training is needed.

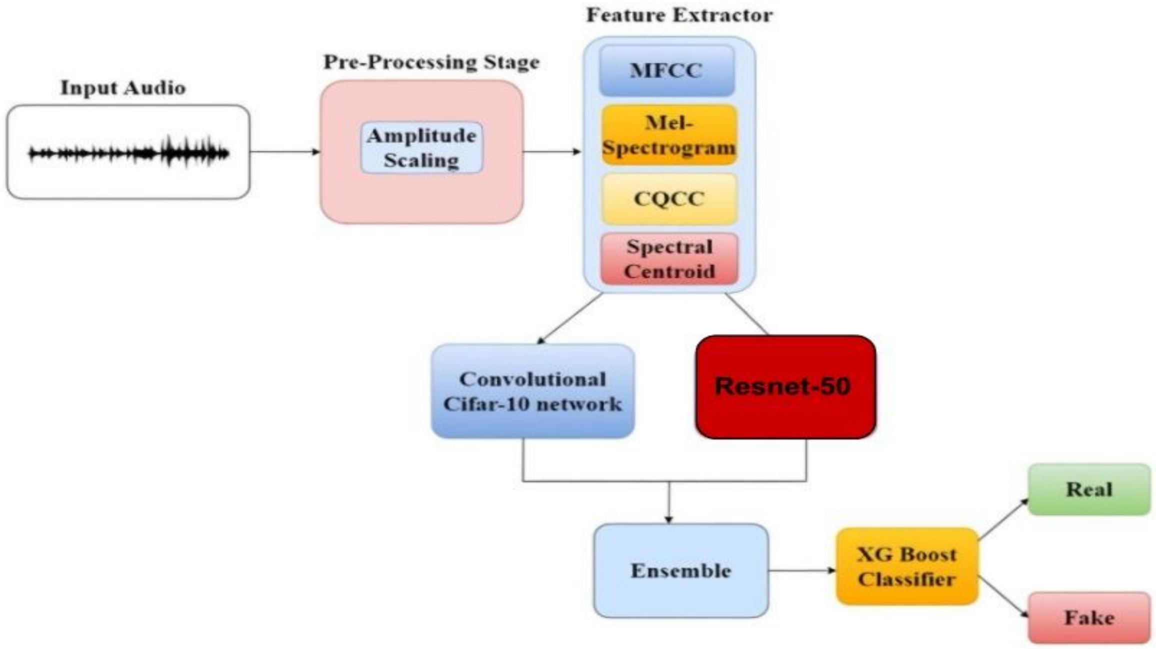

Fig. 2: Framework of proposed audio deepfake detection model

3.2 Audio Dataset

The proposed model is trained and tested on the Fake or Real (FoR) [24] audio dataset, comprising over 195,000 samples. Audio samples in FoR dataset have been generated using cutting-edge speech synthesis technology. The dataset amalgamates multiple sources obtained from various studies, aimed at training models to detect fraudulent speech effectively. FoR dataset is divided into four subsets— for-norm, for-2sec, for-rerec, for-orig-the first three subsets are utilized for the proposed model, while the last subset is excluded due to the absence of audio modification. Description of various subsections of FoR dataset is as follows.

3.2.1 FOR-NORM

The dataset comprises 69,400 audio samples, which includes duplicate audio signals. To ensure data integrity and mitigate redundancy, an initial de-duplication process is applied, resulting in 53,868 unique audio samples. Subsequently, several pre-processing steps are implemented, including the elimination of duplicate entries, adjustment of sample rates, and standardization of volume and channel count. Modifying the sampling rate ensures uniformity in audio data, promoting compatibility across various processing stages. Additionally, controlling volume and channel count standardizes the audio inputs, thereby enhancing the reliability and comparability of subsequent experiments.

3.2.2 FOR-2 SEC

This subset comprises a training set of 17720 audios and a testing set of 3731 audios. Each audio sample in the dataset maintains a consistent duration of 2 seconds and is equally distributed across target classes (fake/real) and genders. This balanced design ensures impartial training and evaluation of algorithms. Additionally, all audio samples adhere to a uniform sampling rate of 41,000 Hz, fostering consistency and compatibility across the dataset.

3.2.3 FOR- REREC

This subset comprises audio samples with a fixed duration of 2 seconds, and contains a total of 13,268 audio samples, encompassing a variety of genders and different classes (real and fake). To focus on audio signal and extract deeper insights, a trimming process is exploited that provides comprehensive audio signal analysis. Sampling rate is 44100 Hz, which is comparable to the For-2sec audio dataset, ensuring compatibility and facilitating seamless integration.

3.3 Pre-Processing

Raw input audio data undergoes several transformations to prepare it for feature extraction and analysis. An essential step in pre-processing is normalization that scales the audio waveform to ensure consistent magnitudes of the extracted features. Normalization is a common practice in audio processing, with the goal of standardizing the data range, typically between 0 and 1 or -1 and 1, to mitigate biases stemming from variations in input data scales.

3.4 Spectral Features for Audio Analysis

Four spectral audio features- MFCC, Mel-Spectrogram, CQCC, and Spectral Centroid are extracted from raw audio input. An input consists of fake and real audio recordings of FoR dataset. These spectral features are extracted using the Librosa library [29], a robust toolkit for audio analysis. This section presents an in-depth analysis of these chosen audio features, with the goal of understanding their inherent traits and operational capabilities.

3.4.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC feature [31] offers a representation of a voice signal's energy distribution across the frequency domain. It is derived through the discrete cosine transform (DCT), which serves to decorrelate the coefficients obtained after applying the logarithm of Mel-scale filter bank. This process is particularly effective at capturing information pertaining to the lower frequency regions of the voice signal.

3.4.2 Mel-Spectrogram

Analysis of an audio signal undergoes segmentation into windowed segments, each of which then undergoes fast Fourier Transform (FFT) to generate a spectrogram. This spectrogram, known as a Mel spectrogram, offers a time- frequency representation of an audio signal. To achieve this, a logarithmic scale called Mel scale is applied, ensuring an alignment of the frequency scale with human auditory system's sensitivity and maintaining equal perceptual distances between frequencies.

3.4.3 Constant-Q Cepstral Coefficients (CQCC)

This is a time-frequency analysis technique designed to closely align with human auditory perception when applied to speech signals [30]. In

CQCC, frequency bins are represented on a geometric scale determined by constant Q values. To compute CQCC, a constant Q power spectrum is

uniformly sampled, and the resulting logarithmic values undergo a Discrete Cosine Transform (DCT) operation [32].

3.4.4 Spectral Centroid

The spectral centroid is an indicator of central point of the spectrum's energy distribution, highlighting where most of the energy is focused. By computing variance of the spectral centroid, we derive the spectral bandwidth, which quantifies the spread or width of the spectrum.

3.5 Fusion Approach for Audio Deepfake Detection

Four spectral features extracted in the previous phase are initially fed to two independent cutting-edge networks, namely, ResNet50 [33] and CIFAR10 [34] and These classifiers are recognized as a leading solution in the field, and serves as a benchmark for assessing the effectiveness of various vocal features. Afterwards, late fusion is implemented that ensembles the prediction results from these two networks to classify deepfake audios.

4. Experiments and Evaluation

To prove the relevancy of various audio spectral features for deepfake detection, various experiments are performed on different subsets of FoR dataset. Dataset is split into 80% training and 20% testing. Different image processing operations are executed utilizing the OpenCV library and the Librosa library in Python. Proposed model is designed, trained, and tested in Pytorch. For training, NVIDIA P5000 GPU is utilized having 16 GB memory. During training, a learning rate of 0.01 is utilized with 40 epochs with batch size 32.

4.1 Evaluation Parameters

Once a model is built and trained, different parameters are used for evaluation. A binary classification model is used to predict whether audio is deepfaked or not. Deepfaked and Real audios which are identified correctly by the model are counted as True Positives (TP) and True Negatives (TN) respectively. The evaluation measures used are equal error rate (EER), binary accuracy, precision, recall, and F1 score. EER estimates the precise threshold at which the false acceptance rate (FAR) and false rejection rate (FRR) are equal, offering a fair evaluation of the system's aptitude for handling real and fake sounds. Precision and recall are useful when dealing with imbalanced datasets, offering insights into correctly identified deepfake. F1 score combines precision and recall minimizing false positives or false negatives. These parameters are crucial in evaluating the proposed model using various performance metrics, as elaborated in Table 2.

4.2. Experimental Results

In this research work, initially ResNet18 [33] model is implemented with individual audio features to evaluate their deepfake detection performance. In all the experiments, four spectral audio feature descriptors- MFCC, Mel-spectrogram, CQCC, and Spectral Centroids are utilized to extract pertinent features from input audio data (FoR). Exploiting different feature descriptors help to analyse and compare their accuracy in extracting useful audio information. The outcomes gathered are presented in Table 3 illustrating performance results on ResNet18 model with respect to individual spectral feature descriptors. After analysis, it is found that MFCC feature yields better accuracy (79.12%), precision (67.54%), recall (68.44%), and F-score (67.80%) in comparison to the others spectral features, when FoR dataset is trained on ResNet18 model.

Table 2: Evaluation Performance Metrics (FPR: False Positive Rate, TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative)

Metric	Delineation	Equation
Accuracy	Evaluation of correct predictions relative to total data instances.	(TP + TN) / (TP + TN + FP + FN)
Precision	Evaluation of accurately predicted positive data relative to all correctly predicted positive data.	TP / (TP + FP)
Recall	Evaluation of accurately predicted positive data relative to all positive data.	TP / (TP + FN)
F-Score	Combines precision and recall, providing a balanced measure.	F-1= $2 * \frac{(Precision * Recall)}{(Precision + Recall)}$
EER	Determines when the false acceptance rate (FAR) and false rejection rate (FRR) are equal.	$FPR = \frac{FP}{TN + FP}$

Table 3: Results on ResNet18 Model for individual spectral feature descriptor

Feature	Accuracy	Precision	Recall	F-score	EER
MFCC	**79.12**	**67.54**	**68.44**	**67.80**	**13.59**
CQCC	72.64	66.34	67.64	59.23	17.23
Mel Spectrogram	68.34	64.55	65.44	65.22	23.86
Spectral Centroid	69.18	62.59	63.75	63.13	20.74

To improve performance of audio deepfake detection, proposed paper ensembles the predictions from CIFAR-10 and ResNet50 networks. Each individual extracted feature is fed and trained on two networks separately: CIFAR-10 and ResNet50. Afterwards, late fusion is employed that ensembles the output from these two independent networks to classify fake or real audios. Table 4 presents the performance results on proposed model with respect to individual spectral feature descriptor. Results of proposed model indicate that fusion of predictions from pre-trained networks yields more accuracy in detecting fake audio in comparison to a single model. Results revealed that proposed model achieves accuracy (99.12%), precision (97.54%), recall (98.44%), and F1 score (98.12%) on MFCC feature, while on CQCC feature, the accuracy (99%), precision (96.34%), recall (98.44%), and F1 score (97.23%).

Additionally, MFCC exhibits a low EER of 0.88, further emphasizing its reliable performance. Although EER with CQCC feature increases slightly to 1.26, the approach remains effective in detecting deepfake audio. It is concluded that Convolution CIFAR-10 network integrated with ResNet50 architecture offers a more complex and deeper network design, capable of capturing intricate patterns and features in audio data.

Table 4: Results on Proposed Model for individual spectral feature descriptor

Feature	Accuracy	Precision	Recall	F-score	EER
MFCC	**99.12**	**97.54**	**98.44**	**98.12**	**0.88**
CQCC	99	96.34	98.44	97.23	1.26
Mel Spectrogram	98.34	94.55	95.44	95.22	1.66
Spectral Centroid	96.15	92.46	93.86	95.20	3.56

Apart from this, performance in terms of EER and accuracy, of eight state-of-the-art audio deepfake detection techniques is evaluated and analysed with the proposed model. Quantitative accuracy-based evaluation of proposed model outperforms eight state-of-the-art audio detection models--DNN, DeepSonar, STN, TCN, SVM, CNN, KNN, and RF. Fig. 3 illustrate EER measure, clearly demonstrating that each detection approach exhibits a unique performance characteristic.

However, effectiveness of a method is greatly influenced by the pre-processing strategy employed. When the FoR dataset is utilized in conjunction with DeepSonar [59], an accuracy of 98.10% is observed as shown in Fig. 4. Conversely, ML based Random Forest (RF) technique exhibited the lowest performance with an accuracy of only 62%. However, proposed model outperforms against the eight models, and achieves an impressive accuracy of 99.12%, 99%, 96.15%, and 98.34% with features MFCC, CQCC, Spectral Centroid, and Mel-Spectrogram respectively. Analysis results indicate that DL methods provide superior performance as compared to ML or the shallow classifiers.

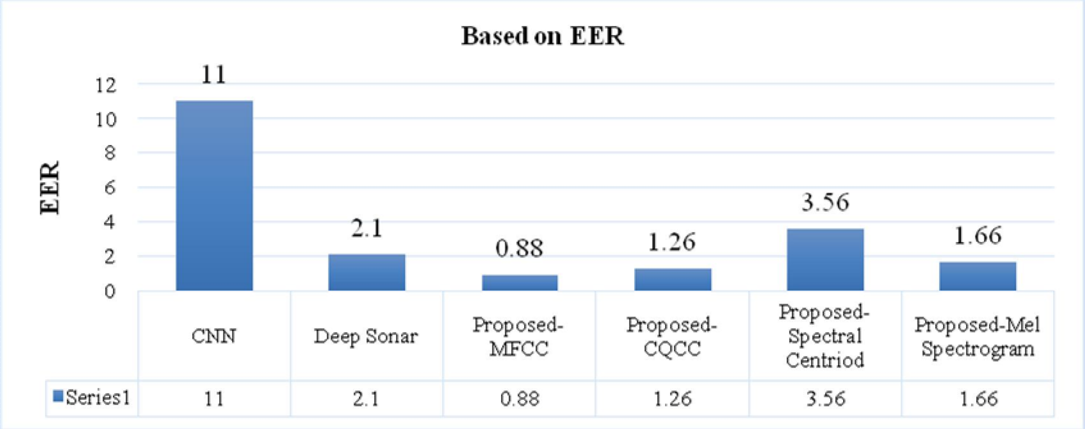

Fig. 3: Quantitative evaluation of EER for various audio detection methods

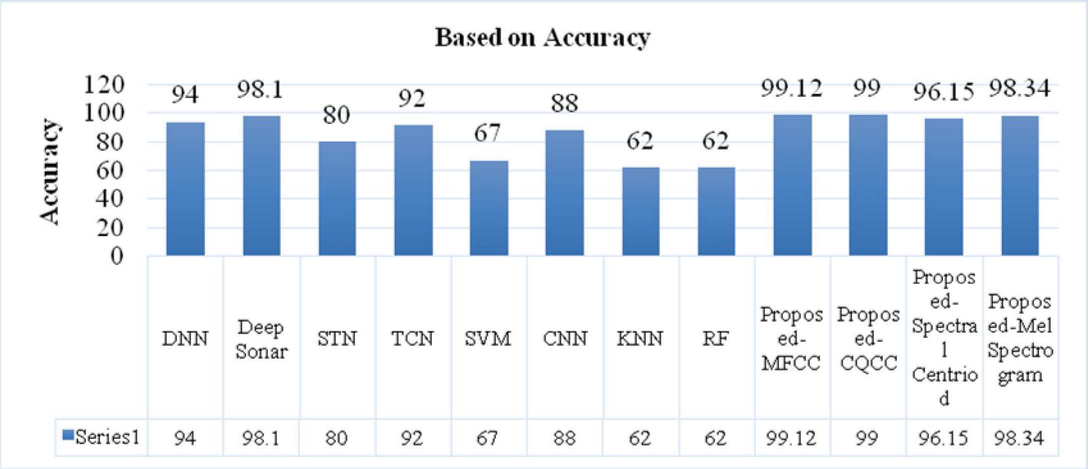

Fig. 4: Accuracy-based quantitative evaluation of different audio detection methods

5. DISCUSSIONS, CONCLUSIONS, AND FUTURE SCOPE

This research paper introduces a novel framework that exploits fusion of predictions from two pre-trained models, demonstrating their significant effectiveness in detecting audio spoofing. It is analyzed that fusion of predictions from pre-trained networks yields more accuracy in detecting fake audio in comparison to a single model**.** Experimental results yielded promising outcomes, with the MFCC feature achieving superior accuracy of 99.12% and a precision of 97.54%. Similarly, the CQCC feature provides an accuracy of 99.00% and a precision of 96.34%. However, the spectral centroid feature showed /somewhat lower effectiveness, with an accuracy of 96.15% and a precision of 92.46%. Several steps contribute to improvement of proposed model. Firstly, normalization step enhances model's ability to learn discriminative features and generalize effectively to unseen audio samples. By normalizing the audio data, variations in amplitude and other factors are reduced, allowing model to focus on the underlying patterns. Secondly, utilization of MFCC, Cepstral, Mel-spectrogram, Spectral Centroid features enable the model to better capture unique characteristics specific to fake audio, resulting in an improved discrimination performance. Additionally, leveraging pre-trained models enhances the model's capacity to extract rich representations and feature hierarchies, potentially improving its capability to identify complex audio patterns. Lastly, utilization of XGBoost boosting algorithm empowers model to handle complex relationships and learn from the extracted features. During training, the model can exploit XGBoost's strengths in capturing non-linear dependencies, leading to accurate predictions. As for future research, one potential avenue is to ensemble the two features using a single model or ensemble of two features with the ensemble of two models, which may lead to more effective results in deepfake audio detection.

References

[1] Y. Wang *et al.*, Tacotron: Towards end-to-end speech synthesis, *Interspeech,* 2017. Available: http://arxiv.org/abs/1703.10135

[2] S. Ö. Arık, J. Chen, K. Peng, W. Ping, and Y. Zhou, Neural voice cloning with a few samples, *Adv Neural Inf Process Syst*, *31*, 2018. Available: https://audiodemos.github.io

[3] S. Kingra, N. Aggarwal, and N. Kaur, LBPNet: Exploiting texture descriptor for deepfake detection, *Forensic Science International: Digital Investigation*, *42,* 2022.
doi: 10.1016/j.fsidi.2022.301452

[4] S. Kingra, N. Aggarwal, and N. Kaur, SiamLBP: Exploiting texture discrepancies for deepfake detection, *Machine Intelligence Techniques for Data Analysis and Signal Processing. Lecture Notes in Electrical Engineering,* Springer, Singapore, 2023, 443–455, https://doi.org/10.1007/978-981-99-0085-5_36

[5] S. Kingra, N. Aggarwal, and N. Kaur, SiamNet: Exploiting source camera noise discrepancies using Siamese Network for Deepfake Detection, *Information Sciences, 645*, 2023,
doi: 10.1016/J.INS.2023.119341.

[6] Y. Zhou and S. L.-P., Joint audio-visual deepfake detection, *IEEE/CVF International Conference on Computer Vision (ICCV),* Canada, 2021. doi**:** 10.1109/ICCV48922.2021.01453

[7] A. Qais, A. Rastogi et al., Deepfake audio detection with neural networks using audio features,
International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), India, 2022.

[8] J. Thies, M. Zollhöfer, and M. Nießner, Deferred neural rendering: Image synthesis using neural textures, *ACM Trans Graph*, *38*(4), 2019.
doi: 10.1145/3306346.3323035.

[9] Vincent J (2018) Jordan peele use ai to make barack obama deliver a psa about fake news.

[10] T. Karras, S. Laine, T. A., A style-based generator architecture for generative adversarial networks, *IEEE transaction on pattern analysis and machine intelligence, 43*(12), 2021.

[11] J. Thies, M. Zollhöfer et al., Face2face: Real-time face capture and reenactment of RGB videos, *Communications of the ACM, 62*(1), 2018, 96-104. https://doi.org/10.1145/3292039

[12] J. Thies, M. Zollhöfer, and M. Nießner, Deferred neural rendering: Image synthesis using neural textures, *ACM Trans Graph*, *38*(4), 2019.
doi: 10.1145/3306346.3323035.

[13] Y. Rodríguez-Ortega, D. M. Ballesteros, and D. Renza, A machine learning model to detect fake voice, *Communications in Computer and Information Science*, *1277*, 2020, 3–13.
doi: 10.1007/978-3-030-61702-8_1/COVER.

[14] D. M. Ballesteros, Y. Rodriguez, and D. Renza, A dataset of histograms of original and fake voice recordings (H-Voice), *Data Brief, 29,* 2020, 105331.
doi: 10.1016/j.dib.2020.105331.

[15] T. Liu, D. Yan, R. Wang, N. Yan, G. C, Identification of fake stereo audio using SVM and CNN, *Information*, *12*(7), 2021.
doi: https://doi.org/10.3390/info12070263

[16] A. Singh and P. Singh, Detection of ai-synthesized speech using cepstral & bispectral statistics, I*EEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR),* Japan, 2021. doi: 10.1109/MIPR51284.2021.00076.

[17] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, Synthetic speech detection

through short-term and long-term prediction traces, *EURASIP J Inf Secur*, *2021*(1), 2021.

[18] M. Todisco *et al.*, ASVspoof 2019 Future horizons in spoofed and fake audio detection, *arxiv.org*, doi: 10.7488/ds/1994.

[19] D. M. Ballesteros, Y. Rodriguez-Ortega, D. Renza and G. Arce, Deep4SNet: deep learning for fake speech classification, *Expert Syst. Appl.*, 184, 2021.

[22] M. Lataifeh, A. E.-D., Ar-DAD: Arabic diversified audio dataset, *Data in Brief, 33*, 2020. https://doi.org/10.1016/j.dib.2020.106503

[23] S. Camacho, D. M. Ballesteros, and D. Renza, Fake speech recognition using deep learning, *Communications in Computer and Information Science*, *1431*, 2021, 38–48.
doi: 10.1007/978-3-030-86702-7_4.

[24] R. Reimao and V. Tzerpos, FoR: A dataset for synthetic speech detection, *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Romania, 2019. doi: 10.1109/SPED.2019.8906599.

[25] P. Aravind, U. Nechiyil, N. P., Audio spoofing verification using deep convolutional neural networks by transfer learning. Available: https://arxiv.org/abs/2008.03464

[26] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, A deep learning framework for audio deepfake detection, *Arab J Sci Eng*, *47*(3), 2022, pp. 3447–3458. doi: 10.1007/S13369-021-06297-W.

[27] S. Kingra, N. Aggarwal, and N. Kaur, Emergence of deepfakes and video tampering detection approaches: A survey, *Multimed Tools Appl*, 2022.
doi: 10.1007/s11042-022-13100-x

[28] A. Dixit, N. Kaur, S. Kingra, Review of audio deepfake detection techniques: Issues and prospects," *Wiley Online Library*, 2023, doi: 10.1111/exsy.13322.

[29] B. McFee, C. Raffel, D. Liang, D. E.-P., Librosa: Audio and music signal analysis in python, *Proc. of the 14th python in science conf.*, 2015. Available: https://www.academia.edu/download/40296500/librosa.pdf

[20] E. R. Bartusiak and E. J. Delp, Frequency domain-based detection of generated audio, *International Symposium on Electronic Imaging Science and Technology*, *2021* (4), 2022.
doi: 10.2352/ISSN.2470-1173.2021.4.MWSF-273.

[21] M. Lataifeh, A. M. Elnagar, I. Shahin, A. B. Nassif, and A. Elnaga, Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations, *Neurocomputing, 418*, 2020.
https://doi.org/10.1016/j.neucom.2020.07.099

[30] M. Todisco, H. Delgado, N. E.-C. S., Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification, *Computer speech and language,45, 2017.*
https://doi.org/10.1016/j.csl.2017.01.001

[31] K. Ito, L. Johnson, The lj speech dataset
Available:
https://scholar.google.com/scholar?hl=en&amp;as_sdt=0%2C5&amp;q=K.+Ito%2C+L.+Johnson%2C+The+lj+speech+dataset%2C+https%3A%2F%2Fkeithito.com%2FLJ-Speech-Dataset%2F%2C+2017.&amp;btnG=

[32] Z. Oo et al., "Replay attack detection with auditory filter-based relative phase features," EURASIP J Audio Speech Music Process, vol. 2019, no. 1, Dec. 2019, doi: 10.1186/S13636-019-0151-2.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[34] A. Karpathy, Convnetjs: deep learning in your browser/