Deepfake Audio Detection with Neural Networks using Audio Features

Abu Qais
Department of Electronics and Communication
Galgotias College of Engineering and Technology
Greater Noida, India
abuqaiselegant@gmail.com

Akshar Rastogi
Department of Electronics and communication
Galgotias College of Engineering and Technology
Greater Noida, India
arastogi23@hotmail.com

Akash Saxena
Department of Electronics and communication
Galgotias College of Engineering and Technology
Greater Noida, India
iiakash59@gmail.com

Arpit Rana
Department of Electronics and communication
Galgotias College of Engineering and Technology
Greater Noida, India
arpitrana9876@gmail.com

Deependra Sinha
Department of Electronics and communication
Galgotias College of Engineering and Technology
Greater Noida, India
dee.sinha@galgotiacollege.edu

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) | 978-1-6654-7258-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCSP53532.2022.9862519

Abstract— **In this paper, a speech spoofing detection system based on Convolutional neural networks using different audio features has been proposed to classify the human speech and synthetic voice, Worst-case scenarios can develop using deepfake audios as threat to assets and image of a person, it can also become a threat to the whole country by unethical uses intended for loss of other party. Using a small voice clip of a person an attacker can develop similar voices. Every audio signal can be represented on a 2D graph plotted by mathematical calculations. The processing of audios into CNN requires a lot of computation, to make a system that can detect deepfake voices with much less computation by conversion of audios to images of audio features (Spectrogram, MFCC, FFT, STFT) and then obtaining the array values as a numeric format which are most suitable to feed. Different approaches for feeding data to model are applied for prediction individually as well as in a concatenated approach.**

Keywords— Convolutional neural network; Audio Processing; FFT: Fast Fourier Transform; MFCC: Mel Frequency Cepstral Coefficient; STFT: Short Time Fourier Transform; Spectrogram; Spoofing detection.

I. Introduction

With advancements in technology, and its increased accessibility to people, So does the deepfakes are also proliferated in media. There were abundance of deepfake videos and audio. Deepfake defines as modified digital media, like photos, videos or audio, in which the image or video of a person, place or thing is replaced with the resemblance of another person or place. In reality, deepfake is one of the most severe concerns confronting modern society. Deepfake has regularly been used to steal the faces of popular Hollywood celebrities over pornographic picture films. Deepfake was also used to generate false information and rumours for politicians.

The process of identifying spoken words by humans and converting them to a readable machine format such as text or command is known as automatic speech recognition. Now people can operate and make their gadgets perform with their voice with the help of this technology. Because of its vast range of applications in many spheres of life, the field of voice recognition has evolved over the last several decades [1], with the major uses of the technology being call centres, dictation solutions and assistive applications, as well as mobile and embedded devices. Due to the huge differences in speech, most automatic speech recognition systems extract features from the acoustic signal rather than using the complete speech signal. These variations include the pitch and speed of the voice, background noise, moods, and expressions. Interacting with a computer voice, keyboard, mouse, touchpad, and so forth is beneficial for persons who have difficulties working with standard interfaces. The process of translating a speech signal into words or phonemes is known as speech recognition [2]. The primary goal of ASR is to overcome all challenges encountered in speech recognition, such as diverse speech patterns, fuzzy background noise, and so on [3, 4]. Deep learning algorithms are used in modern voice recognition systems [5,6]. They are used to represent and model features.

In this study, we provide a fake audio Detection on Convolutional neural networks with Mel-frequency cepstral coefficients, spectrogram, STFT (Short Time Fourier Transform). These different features are used to extract the distinct characteristics feature of each voice signal, which were then used to train the CNN algorithm. The MFCC (Mel Frequency Cepstral Coefficient) technique helps to minimize model complexity and produce improved recognition accuracy.

II. Audio Features

A. Spectral Centroid

The spectral centroid is the main weighted part of the magnitude spectrum. Spectral centroid can be defined as the frequency band where most of the energy is saturated. This corresponds to an audio tone quality (energy, open, dull) as brightness of sound. The spectral spread comprises region around the deviation of spectral centre. Spectral bandwidth plays a vital role with how tone is perceived. The dissipation of energy over a frequency range is proportional to its bandwidth. According to mathematics, it is the weighted average of the distances between a frequency band and the centre of a spectrum.

B. Spectrogram

A spectrogram is a visual depiction of an audio signal's frequency spectrum as it changes over time. As a result, the signal's time and frequency are taken into account. A short-time Fourier transform is used to convert the signal in

978-1-6654-7258-6/22/$31.00 ©2022 IEEE

frequency domain. The STFT of a signal, in its simplest terms, is defined by locally applying a Fast Fourier Transform (FFT) to a small time segment of the signal. Short term Fourier transform is one the simplest features to be considered for transformation, abbreviated as STFT [12]. STFT is used to calculate other parameters such as spectral centroid, spectral slope, and so on.[5] A colour map known as a spectrogram may be used to observe this. With this, an audio track may be described as a discrete time signal x(t), which is a signal in the time domain. Using fixed length segmentation, input signal is splitted into overlapped frames. These frames are equal in length of a specific time window Twin. The no. of sample in a specific window Nwin, calculated as where Fs is sampling rate of a signal.

$$\text{Nwin} = \text{Fs x Twin} \tag{1}$$

Total number of overlapping samples is determined by the hop size parameter ħ. As a consequence, the number of overlapped frames Nb which splits the input signal is calculated as :

$$N_b = \left[\frac{N - N_{win}}{h}\right] + 1 \tag{2}$$

When the discrete Fourier transform (DFT), Xn is computed, the sign is believed to be periodic in each body. On the opposite hand, Xn isn't always periodic because of its finite body length. As a result, the spectral power of a given frequency spreads or leaks into adjoining frequencies. Spectral leakage is the time period for this. A window feature is a multiplicative feature carried out to every pattern in a body to lessen spectral leakage by making the sign periodic [12] presents an entire evaluation of many forms of window functions, where α = 0.54, β = 1 - α|12|.

$$W(t) = \alpha - \beta Cos\left(\frac{2\pi t}{N_{win} - 1}\right) \tag{3}$$

C. Mel Spectrogram

People seem to hear sounds logarithmically. Low frequency differences are more noticeable than high frequency deviations. We can easily distinguish between 500Hz and 1000 Hz, but it will be difficult to distinguish between 10000 Hz and 10500 Hz, even though the two pairs are the same distance. As a result, a fine scale was created. This is a logarithmic scale, meaning that equal lengths of scales correspond to equal sensing distances. Mel's Scale, based on human hearing, is a scale based on perception.

As a result, a mel spectrogram is simply a spectrogram having melscale frequencies. The spectrogram frequencies are compressed with mel scale formula to get result as mel spectrogram frequencies. Mel's Scale, based on human hearing, is a scale based on perception. The mel value can be determined by the following formula [13] for the frequency f(Hz):

$$mel(f) = 1127.01048 log\left(1 + \frac{f}{700}\right) \tag{4}$$

The below figure 1 is a spectrogram of Human Voice. It can be seen that the human voice is not evenly distributed in the lower frequency over time.

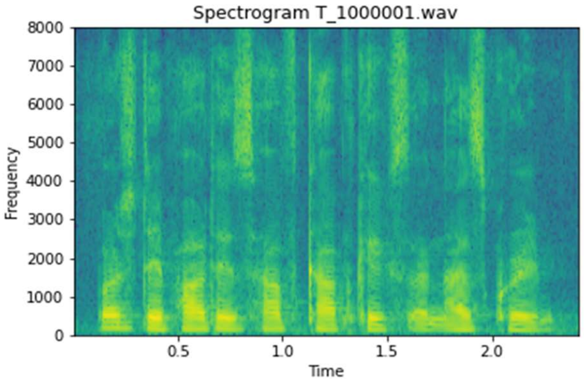

Fig 1: Spectrogram of human voice

The below Image figure 2 is the spectrogram of a synthetic voice and is clearly visible that synthetic voices have a more even distribution in lower frequencies compared to human voice.

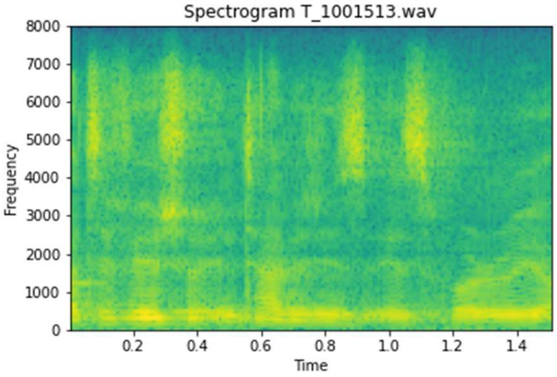

Fig 2: Spectrogram of synthetic voice

D. MFCC

A signal's cepstrum comprises data regarding the rate of change in its spectral bands. A cepstrum is just the spectrum of the log of the temporal signal. The ensuing spectrum, that is neither in the frequency region nor in the time region, is known as the quefrency are the coefficients that make up the mel-frequency cepstrum. The signal cepstrum stores information about the spectral range's rate of change. The cepstrum communicates the many values that make up a sound's formants (a distinguishing component of its quality) and timbre. As a result, MFCCs are beneficial in deep learning models.

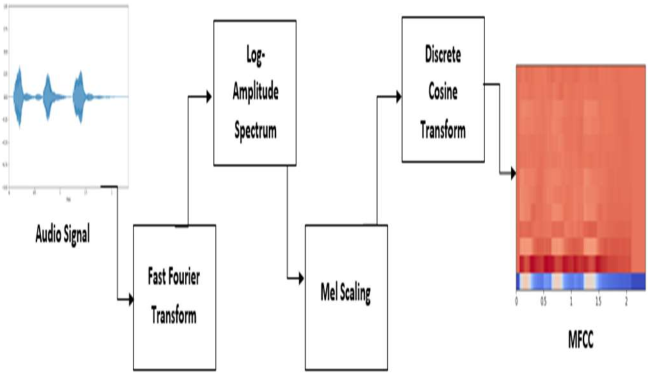

Fig 3 : Steps to extract MFCCs from an audio signal.

Cepstrum is just the logarithmic spectrum of a signal's temporal spectrum. The frequency domain refers to the spectrum that is neither in the frequency domain nor in the time domain. Figure 3 shows how the MFCC feature is extracted from audio signal. MFCC coefficients (Mel Frequency Control Coefficients) make up the ultra-low frequency cepstrum.

The below figure 4 and figure 5 is a MFCC feature extracted from Human Voice as well as Synthetic voice. It can be seen that the human voice is concentrated in the lower frequency over lower magnitudes.

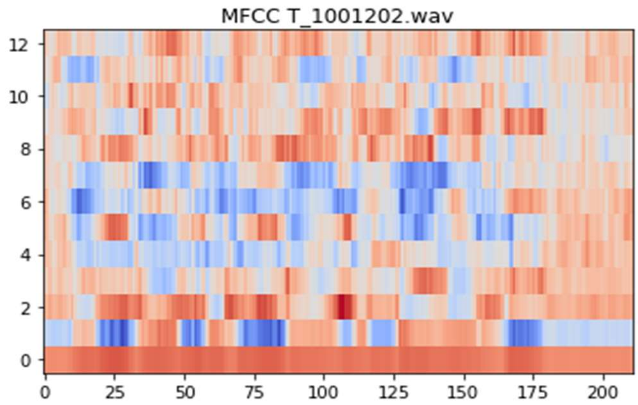

Fig 4: Human Voice MFCC

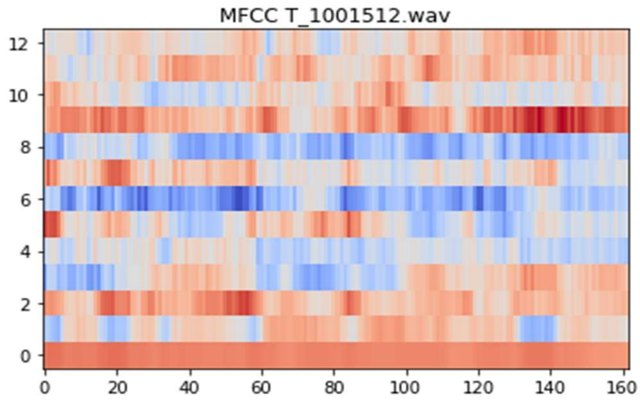

Fig 5: Synthetic Voice MFCC

E. *FFT (Fast Fourier Transform)*

Fast Fourier Transform (FFT) is an algorithm which is used to obtain the Discrete Fourier Transform (DFT) of a sequence in our case sequences are contained within audio variables, it also computes the Inverse Discrete Fourier Transform (IDFT) of the sequence. It converts the audio into individual spectral components which provides us the information of frequencies of a signal.[11] A signal gets sampled over period of time and obtained as frequency components. FFT converts from its original dimensions such as time to obtain frequency presentation by decomposing signal into different frequencies of the signal. FFT quickly integrates such changes by making the DFT matrix a product of small (especially zero) objects.

DFT obtains the spectrum of signal which very useful in the fields of signal processing. The DFT or FFT algorithm can convert this unique time zone signal into a frequency zone.

The Fourier transform (FT) of the function f(x) is the function F(ω), where

$$F(w) = \int_{-\infty}^{\infty} f(x)e^{-iwx}dx \quad (5)$$

and the inverse Fourier transform is:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(x)e^{iwx}dw \quad (6)$$

The below figure 6 is a FFT of Human Voice. It can be seen that the human voice is concentrated in the lower frequency over lower magnitudes.

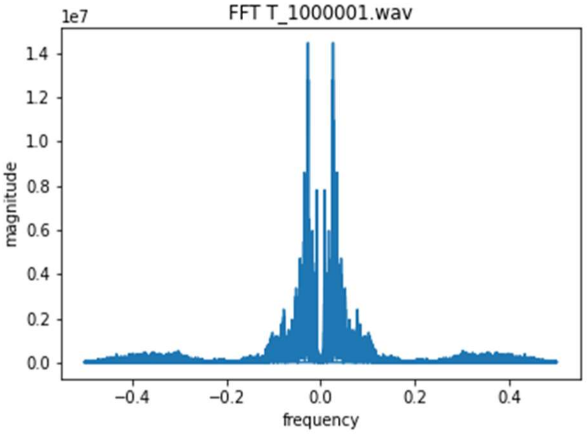

Fig 6: Human Voice FFT

The below figure 7 is a FFT of Synthetic Voice. It can be seen that the synthetic voice is not concentrated in the lower frequency over lower magnitudes.

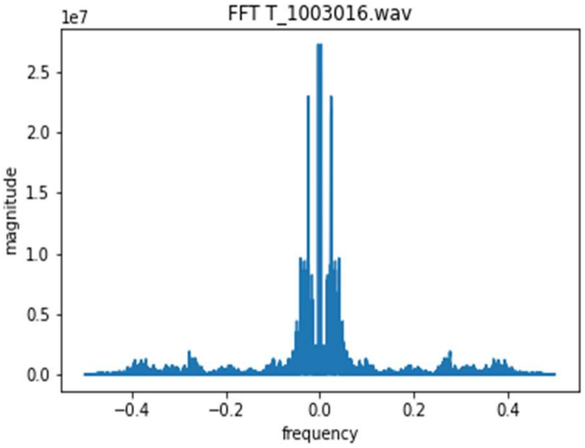

Fig 7: Synthetic Voice FFT

F. *STFT (Short Term Fourier Transform)*

Short Term Fourier transform is an extension of FFT which s a sequence of window transformed signal. STFT provides frequency information when frequency components vary over time. They also provide the basis for sound signal processing functions intended to mimic a person's vision, such as the sensitivity of a hearing area. The FTFT of a signal is given as

$$STFT\ \{x(t)\} = X(T,W) = \int_{-\infty}^{\infty} x(t)w(t-T)e^{-jwt}dt \quad (7)$$

where, t is *time*
x = *input signal*
w = *length of window*

Fourier short-term transformer (STFT), is used to determine the sinusoidal frequency and content of local signal components as it changes over time. In fact, the process of assembling computer STFTs is to divide the long-term signal into shorter parts of equal length and to calculate the Fourier variance for each short-term segment. This reveals the Fourier spectrum in each short section.

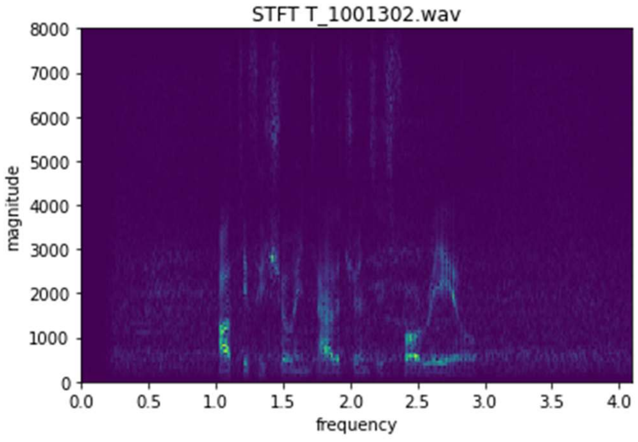

Fig 8: Human Voice STFT

The above figure 8 is a STFT of Human Voice. It can be seen that the human voice frequencies are distributed over and have higher frequencies with larger magnitudes.

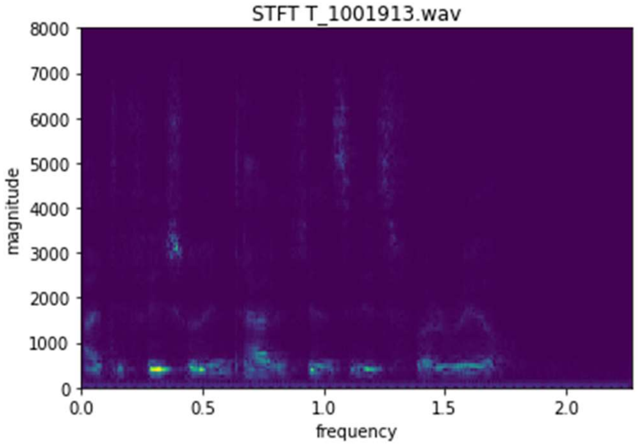

Fig 9: Synthetic Voice STFT

The above figure 9 is a STFT of Synthetic Voice. It can be seen that the synthetic voice frequencies are distributed in lower frequencies and also have lower magnitude values.

III. System Model

The audio signal, which is characterized as a series of raw audio frames or feature vectors created by humans. CNNs work by convoluting input with learnable kernels. A 1-d temporal convolution is used in the place of an array input characteristics. 2-Dimensional time-frequency convolution is commonly employed as an alternative. A time-domain 1-d convolution is utilized for raw data inputs with waveforms. A convolutional layer, in most circumstances, generates many feature maps (channels), each of which corresponds to its own kernel. There are RNNs can experience vanishing/exploding gradients. To solve this, a number of strategies have been developed. The thick layers can be eliminated for sequence labelling to create a fully convolutional network. RNNs are artificial neural networks in which nodes' connections build a directed or undirected graph over time. RNNs use a unique approach to modelling sequences. Sound segmentation is a technique for separating a desired signal from a mix of sounds for subsequent processing. The images of different features such as spectrogram, STFT, FFT, MFCC are feeded to CNN model with defined parameter directly as well as in array form. They automatically simulates the inputs temporal dependence and allows the receptive field to stretch endlessly into the past.

A. CNN (Convoltutional Neural Networks)

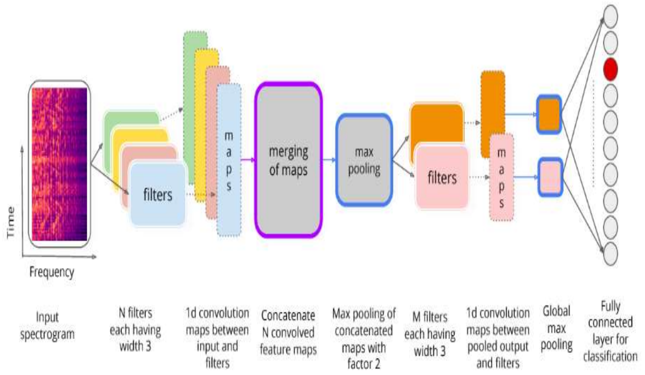
Time
Frequency
filters
m a p s
merging of maps
max pooling
filters
m a p s
Input spectrogram
N filters each having width 3
1d convolution maps between input and filters
Concatenate N convolved feature maps
Max pooling of concatenated maps with factor 2
M filters each having width 3
1d convolution maps between pooled output and filters
Global max pooling
Fully connected layer for classification

Fig 10: CNN Model with Audio feature image as input

A CNNs work by convoluting input with learnable kernels. A 1-d temporal convolution is used as input of spectral features, a 2-d time frequency convolution is used for raw data and time dimensional time-frequency convolution is frequently used for raw data, and a time-domain one dimensional convolution is used inputs with waveforms. Typically, a convolutional layer has the role to compute several feature maps (channels), each corresponding to its own kernel. Above these convolutional layers, there are pooling layers. The learned feature maps can be down sampled using layers. CNN is made up of several convolutional layers. interspersed by pooling layers, then one or more dense layers. The dense layers can be eliminated for sequence labelling. in order to create a fully convolutional network (FCN).The receptive field (the number of samples or spectra) is a term that refers to the number of samples or involved in the computation of a CNN prediction). Due to above facts, the model of a CNN is now mostly decided through observation and analysis depending upon a validation error, leading to various rules to follow such as fewer layers parameters for less data [10], In consecutive convolutional neural networks, the size of feature maps gets smaller layers, these are taking into account the required size for appropriate analysis.

B. ReLu (Activation function)

ReLu stands for Rectified Linear Unit. It is an activation function. An activation function plays the role of transforming the summed weight from a node to the output of that node. This value is known as summed activation of that node. A ReLu looks like a linear function but acts as a non-linear function. The function returns 0 if it receives any negative input, but for any positive value x it returns that value back. The ReLu function can be mathematically represented as :

$$f(x)=max(0,x) \tag{8}$$

This *max()* function returns the greatest value between 0 and *x*. If the value is less than 0 then it will return the input as the output of the node. ReLu allow model to account for non-linearities and interactions. ReLu is a widely used activation function.

C. Dataset

The ASVspoof 2017 was used to train and test the system [21]. In all of our experiments, we used the ASVspoof 2017 database as described this database is intended to explore a variety of activities related to verification, from detection of deception to counter measures play the attack again. For this reason, we have considered only part of it of the database associated with the acquisition of a transaction a problem considered in our work, which is described as logical access to the database. The data is split into training and testing in the ratio of 80:20. Methods and Technique. The main challenge faced in any model is the computational resources. A model is irrelevant if it fails to perform on a large scale or becomes in economical in the long range. As our paper deals with the handling of audio data which is by-default large size comparing to images and numerical data. To overcome computational restraints, we converted the data formats. The audio is collected as an '.wav file' which is a format for storing waveforms. Images are small in size as compared to audio. These audio files are transformed into images. The above explained concepts of audio features such as MFCC, FFT, STFT, SPECTROGRAM are performed on data which is feed into our model. All the features have a same source of generation i.e. from waves of sound.

D. Model Parameters

Table 1 describe the parameters used in CNN model.

Table: 1

Parameters	**Description**
Epochs	30
Steps per Epoch	8
MaxPooling2D output shape	(nan, 99, 99, 16)
Convolutional Layer Filters	16
Audio Features Input Dimension	200x200x3
Max Pooling Layer Output Dimension	100x100x3

IV. Results

We described a synthetic speech detection method in this research to prevent spoofing attacks on biometric speaker verification systems that generate the impostor signal via synthetic voice adaption or conversion. For synthetic signal feature extraction, we investigated MFCC, STFT, FFT, Spectrogram parameterization and other audio features. Because synthetic speech signals are nonlinear and unpredictable, so different non linear features are used. The model is implemented on python using Keras Sequential Convolutional Neural network on a connected host time of Google Colab. The Activation function used is 'ReLu' For feature extraction we have used librosa and scipy: wavfile, signal.

Table: 2

MODEL	**ACCURACY**
df_MFCC_relu	75.97%
df_STFT_relu	73.34%
df_FFT_relu	71.48%
df_SPEC_relu	70.86%
df_concated_relu	80.47%

The above table 2 is the accuracy table for Convolutional Neural Networks with ReLu as an activation function prediction. The df_MFCC_relu means MFCC features with Relu as an activation function give 75.97% accuracy to classify the human speech with synthetic voice. The spectrogram here also performed the least with an accuracy of 68.9% due to the fact that the training data was smaller compared to other data and the concatenated data which had all the features combined performed the best. Here it can be seen that MFCCs outperformed every other feature as the MFCCs have the most vibrant images which means more pixel values of our arrays are not the maxima and minima(255, 0) which gave the neural network to learn more on this data. The below bar plot is the comparison of accuracy scores.

For Validation purpose the Receiver Operating Characteristics Curve (ROC) as shown in figure 11, is preferred over accuracy as an ROC tells more than just prediction. It is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate(TPR) and False Positive Rate(FPR).

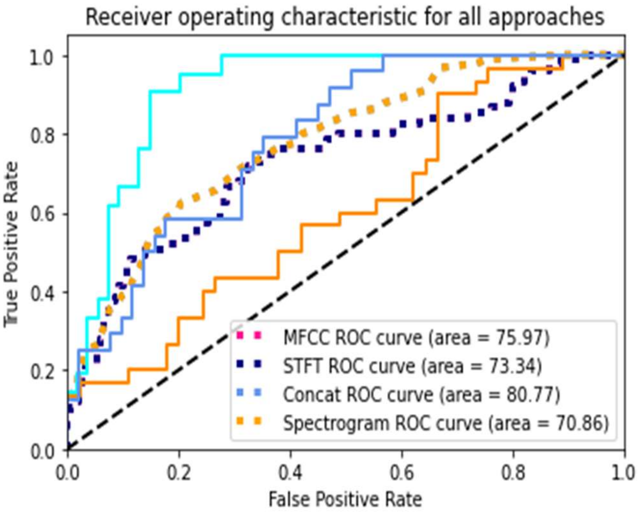

Fig 11: ROC curve area for all approaches.

CONCLUSIONS

In In this proposed research the main motive was to analyze and predict sound waves considering computational constraints. The audio files were converted into images to obtain arrays that were fed to the Convolutional Neural Network (CNN). We by-passed the feeding of audio to predict, yet we used to convert audio into images of audio features and then converted those images to an array which gave us a numerical data input. The all numerically obtained data helped us avoid the errors which were about to phased due to unnecessary and unavoidable noise. The studied literature review also gave the abstract that the MFCCs perform best compared to other features and we proved the same in our research work. Both models performed least for df_SPEC, which was the data frame of arrays of Spectrograms. The most efficient approach was concatenation of all audio features in the Convolutional Neural Network using rectified linear unit activation function (ReLu) activation function CNN model on all arrays features combined data. The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. The strategy to obtain the best result from different combinations of features help us obtain the accuracy as well as understanding of dependencies of features upon audio composition. Overall we found that MFCC features performed the best in all individual scenarios all of which were dominated with concatenated approach.

The dataset was few years old it would be so better if an updated dataset will be used as the spoofing technology is evolving faster than detection technology. Furthermore Recurrent Neural Networks (RNN) can be preferred over Convolutional Neural Networks as RNNs feed the results back into the network working as a feedback. Graph Neural Networks (GNN) also overcome the limitation of CNNs underperformance in non-Euclidian space.

ACKNOWLEDGEMENT

We would like to express our gratitude to everyone who assisted us in completing our project " Deepfake Audio Detection with Neural Networks using Audio Features " Mr. Deependra Sinha, who is an AI domain specialist mentored us throughout the research, deserves special thanks for his invaluable assistance. His presence as a mentor has been incredibly encouraging. We would like to thank Gaurav Mehra for his generosity and knowledge, which has benefited this project in several ways and spared us from numerous blunders. We would like to express our gratitude to Dr. Lakshmanan, the chair of the Department of Electronics and Communication Engineering, for allowing us to participate in this exciting initiative.

REFERENCES

[1] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206-219.

[2] Xie, D., Zhang, L., & Bai, L. (2017). Deep learning in visual computing and signal processing. *Applied Computational Intelligence and Soft Computing*, *2017*.

[3] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, *7*, 19143-19165.

[4] Almars, A. M. (2021). Deepfakes detection techniques using deep learning: a survey. *Journal of Computer and Communications*, *9*(5), 20-35.

[5] Subbarao, M. V., Padavala, A. K., & Harika, K. D. (2022). Performance Analysis of Speech Command Recognition Using Support Vector Machine Classifiers. In *Communication and Control for Robotic Systems* (pp. 313-325). Springer, Singapore.

[6] Balamurali, B. T., Lin, K. E., Lui, S., Chen, J. M., & Herremans, D. (2019). Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access*, *7*, 84229-84241.

[7] Shafiee, E., Mosavi, M. R., & Moazedi, M. (2018). Detection of spoofing attack using machine learning based on multi-layer neural network in single-frequency GPS receivers. *The Journal of Navigation*, *71*(1), 169-188.

[8] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, October). Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In Proceedings of the 28th ACM international conference on multimedia (pp. 2823-2832).

[9] Al-Dulaimi, A. W., Moon, T. K., & Gunther, J. H. (2021). Voice Transformation Using Two-Level Dynamic Warping and Neural Networks. *Signals*, *2*(3), 456-474.

[10] Gomez-Alanis, A., Peinado, A. M., Gonzalez, J. A., & Gomez, A. M. (2019, September). A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In *Proc. Interspeech* (Vol. 2019, pp. 1068-1072).

[11] Liu, T., Yan, D., Wang, R., Yan, N., & Chen, G. (2021). Identification of Fake Stereo Audio Using SVM and CNN. *Information*, *12*(7), 263.

[12] Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., & Černocký, J. (2019). Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge. *arXiv preprint arXiv:1907.12908*.

[13] Lou, J., Xu, Z., Zuo, D., & Liu, H. (2021). Feature Extraction Method for Hidden Information in Audio Streams Based on HM-EMD. *Security and Communication Networks*, *2021*.

[14] Bhangale, K. B., Titare, P., Pawar, R., & Bhavsar, S. (2018). Synthetic speech spoofing detection using MFCC and radial basis function SVM. *IOSR J. Eng.(IOSRJEN)*, *8*(6), 55-62.

[15] Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. *arXiv preprint arXiv:1001.2267*.

[16] Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. (2019). End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*.

[17] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97.

[18] Sarfjoo, S. S., Wang, X., Henter, G. E., Lorenzo-Trueba, J., Takaki, S., & Yamagishi, J. (2019). Transformation of low-quality device-recorded speech to high-quality speech using improved SEGAN model. *arXiv preprint arXiv:1911.03952*.

[19] Punnappurath, A., & Brown, M. S. (2019). Learning raw image reconstruction-aware deep image compressors. *IEEE transactions on pattern analysis and machine intelligence*, *42*(4), 1013-1019.

[20] Lyu, S. (2018). Detecting deepfake videos in the blink of an eye. *The Conversation*, *29*.

[21] ASVspoof 2017: Automatic Speaker Verification Spoofing And Counter measures Challenge dataset