Deepfake Audio Generation using LSTM-GAN and Detection using CNN Classifier

Sugandha Sharma
Master of Science – Data Analytics
National College of Ireland
Dublin, Ireland
x21236577@student.ncirl.ie

Sakshi Kishor Khanvilkar
Master of Science – Data Analytics
National College of Ireland)
Dublin, Ireland
x22117776@student.ncirl.ie

Gurpreet Kaur Bhuie
Master of Science – Data Analytics
National College of Ireland
Dublin, Ireland
x21231061@student.ncirl.ie

***Abstract*—The process of generating and altering recorded audio has become very easy with the introduction of deepfake technology. This technology brings benefits in domain like entertainment or aiding people who are deaf. Along with benefit it also brings obstacles in security, where it is difficult to identify the real or fake audio. This study tries to find a solution to this problem by building a model that can differentiate between the real and fake audio. Two separate models are built to achieve this objective, first is a generative adversarial network (GAN) that generates fake audio and another is a classifier that distinguishes between real and fake audio. The generative adversarial network has two parts: generator and discriminator. The discriminator in GAN tries to distinguish the real data into real and generated data as not real, whereas generator arises to the point where the discriminator can tell differences between real and fake data and improve their code accordingly. GAN model is built by stacking these models which used LSTM and Dense layer to process audio data which is sequential in nature. The classifier model is built using CNN and is trained on a series of real and fake data to effectively classify real data from fake data. This paper also highlights the challenges faced during study.**

***Index Terms*—Generative Adversarial Network, tensor flow, Long-Short Term Memory network, Deep Learning, CNN, Classifier.**

I. Introduction

In recent year deep learning has moved from hand-crafted aspects that needed detailed knowledge to feature learnt from new audio like short-term Fourier transform. This leads to developing models that require less knowledge, still at the cost of data, computational power, and training time. For example, deep autoregressive technique is directly implemented on raw data also on Mel-scaled spectrograms [1]. These models take more time to train a conventional GPU also their generative procedure is slow. Whereas GAN obtains similar audio mixing quantity and quicker generation time. Nowadays the development of smart phone cameras and easily accessible to internet and social media platform have made uploading audio-based video easier. Like many other problems in growing technology has given rise to new problems. The most advanced way of handling the audio and video sample is called "Deepfake" [2]. The necessity for accurate and consistent audio analysis and identifying systems in various applications leads to the development of Generative adversarial Network (GANs) for audio classification.

Looking at the example of detecting fraud in customer service call center where it is important to check the genuineness of audio transaction, like validating a caller's identity during a financial transaction, is mandatory for maintaining trust and security. Also, huge amount of speech recording is shown over the internet every day and detecting false on it is difficult [3]. On other hand, the different and large amount of fraudulent data and real audio transactions makes it difficult to build a classifier model. Existing Techniques constantly fail to analysis the minute detailed variation in fraudulent data; hence this leads to misclassification and exposure to new fraud strategies.

Generative adversarial networks provide a convincing solution. By training a GAN on a small amount of real and fake data, the generator of the network can produce extremely credible fake audio samples. Meanwhile, the discriminator can learn to distinguish between real and fake audio. The generate trials efficiently expand the dataset giving wide range of fake audio and variation, further combining with the real audio this expanded dataset will be able to develop more strong fraud detection model. Therefore, The GAN- classifier can precisely identify the abnormalities in audio, react to the coming fraud strategies, and decrease the risks connected with financial transaction. Moreover, the audio deepfake attacker not only attacked common people but also politicians and governments[4].

In 2019, fake hired AI-based company generated the voice of CEO and scammed over than USD 243,000 through the phone[1]. By implementing the GAN classifier gives an advantage in the area of audio analysis, can detect the oddity within audio data. This GAN classifier not only detects irregularity but also have the capacity to effectively resist new false technique, by reducing risk in financial transactions.

One unique feature of GAN is they can generate realistic audio data and broad range of data sample that nearly approach the target pattern. The complete GAN works in 2 steps on that is generator generate audio that are difficult to differentiate from actual audio, but the discriminator network acquires to differentiate between real and false audio. This adversarial property enhances the classifier capability to identify small

[1] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrin

false that may result in fraud, allowing for more accurate analysis of mistakes

When it comes to audio manufacturing it is necessary to generate exact time resolution of sampled audio data, as GAN is able to generate images with local and global features its exposure to audio sample remains untouched. In such cases WaveGan is the type of GAN used for unsupervised audio sampling. WaveGan can produce one-second part of audio waveform along with maintaining overall consistency to generate sound [5]. Furthermore, there are different deep learning models which can be combined with the GAN model, one of which is LSTM (Long Short -Term Memory). LSTM are a type of RNN (recurrent neural network) that works on modelling data in sequences and capture time-dependences in audio signal. By combining the LSTM into the GAN design, the final model can perfectly capture the complex elements that identify the audio data and increase the capacity of detecting the variance and distinguish between real and fake.

As fraud in various fields is increasing day by day, this combined method provides an effective way to detect and minimize the fraud and protect against audio-based manipulation. For this study will be building the LSMT-GAN model for generation of audio and further it will be classified to identify the fake and real audio. The initial step of this study will start by sorting the input audio followed by implementing functions for generator and discriminator in google colab then the audio generated from GAN will be further classified into real and fake. Overall, we are analyzing the capabilities of generative adversarial networks (GANs) with LSTM in identifying the differences between real and fake audio. Audio classification is involved in pattern recognition and artificial intelligence. Audio data consists of four layers frame, clip, shot, and high level. Three data features are extracted to construct feature vectors and then it classifies the audio data. It differentiates between real and fake data[6].

It is very difficult to identify the authenticity of AI-generated recordings. This can be useful in many fields like media, and security, and for helping people who lost their voices. This motivated us to explore this direction. In our report, we will try to produce synthetic audio clips using the GAN model and do a classification between real and fake audio clips. Hence our research question for this study is

To what extent LSTM can be used to build GAN for fake audio generation?

The reminder of the document is structured in four more section named; Section II as Related Work where we have critically evaluated the research papers of different authors on the GAN model, audio detection and classification model. Section III is Methodology which explains the methodology in detail. It tells what is the problem statement, and dataset and gives step-by-step explanations of the different models used in the report. In section IV Evaluation and Results we have discussed the results interpreted after the implementation of the code. Section V is Conclusions and Future Work where various challenges and future work are discussed for this research along with conclusion.

II. Related Work

There have been many researches done in the field of generative AI and deep fake audio generation and detection. Some of the interesting and motivational studies will be discussed in this section. Social media are one of the most popular way to influence people and recent study sheds light on the problem caused by wide use of social networking platforms for digital video sharing, including the AI-generated deepfake video.[2] The videos which are made by using generative adversarial networks (GAN) techniques can be a threat to social media by spreading false information and spam. The implemented method in this study is to detect deepfake video by concentrating on distinct elements affected during the generation process. For feature extraction and sequential analysis, a combination of CNN (Conventional Neural Network), RNN (Recurrent Neural Network) and LSTM (Long short-term memory) deep learning techniques are used in this research paper. The dataset used is fake and real video data and the accuracy for audio detection is 85% and that for video accuracy 81.45%. The method in the research paper gives confirming results in detecting deepfake video with good accuracy. Moreover, chances of development, especially in real world scenarios where speed and precision are key.

With the advancement in generation of fake data the need for detecting the fake data accurately and precisely is also increasing. A recent study[7] proposes a solution for improving machine learning classification task using actual audio data. This paper focuses on two parts first is by using a generative AI for generating new audio data and second by using learning method to generate audio with specific features that are not present in original dataset to enhance the classification performance. These features are examined further with SVM (Support Vector Machine), on SVM trained on the original dataset the unweighted average recall is 75%, the SVM trained on the regulated dataset the unweighted average recall is 78.8%. The paper mainly focused on soundscape classification limiting the proposed technique use to particular setting which could resist to adapt other audio classification task.

Another study [1] represents a different audio signal for Generative Adversarial network-based audio manufacturing aiming to generate musical sound generation. It uses various techniques like complex valued short time Fourier Transform (STFT), constant- Q transform, Mel spectrogram. The author elaborates about quantitative metrics and qualitative metrics. Hence the result shows that STFT gave better findings. However, the evaluation expose certain limitation which cannot capture non audible distortions.

In one of the study, author talks about the two-stage analysis which has CNN from frame level extraction and RNN for identifying temporary inconsistency in the face swapping process. This one of interesting study in this area sheds light on misuse of the deep fake technology and suggested exploring the system robustness encountered during the training process.[8]

In another study the author has used a multimodal approach to discriminate between real and fake content this gives us motivation to think how deep learning networks can exploit audio, video, and emotional cue. [9] In another study, author has focused on deepfake audio detection using GAN. The auto-encoder approach combined with Mean Squared Error on the ResNet-18 model has been proposed. The study shows and interesting way to bridge gap between GANs and audio, but the limitation is the lack of ground truth latent vectors for real samples. The author has evaluated WaveGAN[10] and suggested exploring more GANs.

Another interesting study[11] highlights the Audio Deepfakes that initially started as an audiobook and now moved towards imitation and synthetic-based deep fakes. For detecting AD Machine learning and Deep learning methods are used. The limitation is that it does not cover accented voices or real-world noises. The study shows that the choice of approach has a greater impact on performance than the audio features used, resulting in a compromise between accuracy and scalability. The two main limitations of this study are overlapping sounds from different construction activities and the inability to determine the start and end time of the activity. CNN is used for feature extraction and RNN for noisy environments. Dual threshold output is used to identify the timing of each activity. The model is pre-trained and validated in a modular construction factory. The main focus of this study is on monitoring construction activities. [12]

Forensic analysis of deep fake audio detection methods in another interesting area i.e., to discriminate between synthetic and real voice. It has presented audio files such as MFCC, Mel-spectrum, Chromogram, and spectrogram representations. The author has employed Deep learning models to analyse features and evaluate their effectiveness. Comparison of architecture and techniques have been done but the main focus is to analyse the authenticity of audio files. [13] The author describes the deepfake detection system across different languages. It focuses on reliable deep fake detection methods. In this study exploration of domain adaptation techniques and adversarial-based paradigm is done to differentiate between real and fake audio [14]. In this paper multiple rounds of evaluation are done in a fake audio game, manipulated region location and deep-fake algorithm. They have used LFCC features with a light convolutional neural network. To determine if a speech signal was produced by a known or unknown deepfake algorithm, Baseline S05 utilized a straightforward method called SoftMax with a thresholding procedure based on probability [15]. This study has given an overview of neural network architecture for simulating audio effects. Wavenet and conditional GANs are used which helped us in our project to generate waveforms, and spectrograms [16]. In this research paper limitations have been highlighted of deep generative models to implement on low-resource devices for audio synthesis.It also focused on the benefits of adversarial periodic feature distillation [17].The paper evaluates sound-based diagnostics using GAN to augment data and address the challenges. Author proposes CNN for diagnosing covid 19 audio data and also explains the ways of improving accuracy[18].

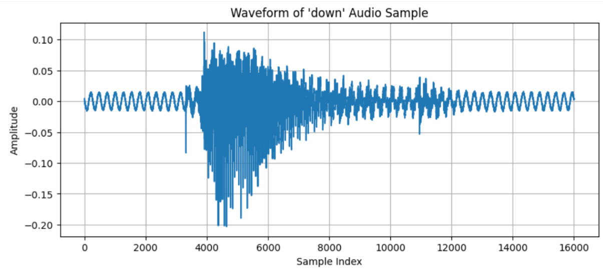


Fig. 1. Caption for your image.

III. Methodology

To address the research question 2 models were developed, a GAN model using LSTM that will generate deepfake audio and classifier model that classify the audio is real or fake. CRISP-DM methodology used to implement these models and the steps used are as follows. Also classification model is used to detect if the audio is real and fake.

A. Problem Statement

The purpose of this research is to generate fake audio using LSTM model and build a classifier to differentiate between fake and real audio. The data used to train GAN model is simple speech commands[2] by google which includes audio of about one second long and a sequence length of 16 kHz. This research is aiming to generate expected audio of similar quality and length. Speech commands available in dataset are: ’up’, ’right’, ’no’, ’yes’, ’left’, ’go’, ’down’, ’stop’.

For the classification model another dataset has been taken which is [3] divided into subfolders as real and fake. The aim is to process the dataset, build the model and classify to discriminate between real and fake audios.

B. Understanding the Data

The data of speech commands which is used to train GAN is diverse in terms of content as the words are spoken by different people. This resulted in data having many different frequencies and modulation for one single world. Data is also examined for any imbalance and missing value, but it was not observed. An audio clip is basically a waveform and can be plotted using a spectogram like in Figure 1 below shows a wave for the word down. There are a total to 8000 records for about 8 different words. The frequency of audio is 16 kHz and for this reason the sequence length set as 16000 is this experiment. Different options of batch sizes were tried during hyper-parameter tuning and it was observed that lower batch size resulted in slower training rate but use up less resources. Though 64 is a comparatively higher batch size but it resulted is fast training and hence was selected for this experiment based on the resources available. Likewise, other parameters that is to be used are also studied and hyperparametres are

[2] Speech Command Dataset: http://storage.googleapis.com/download.tensorflow.org/data/mini_speech_commands.zip

[3] 2 second audios: https://bil.eecs.yorku.ca/datasets/for-2sec.tar.gz

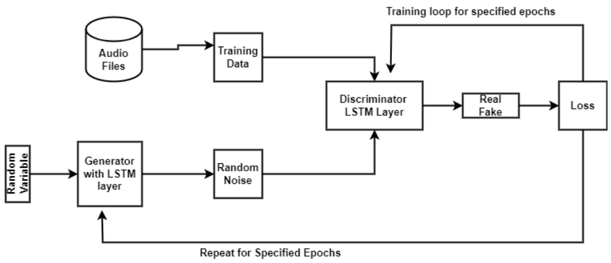


Fig. 2. Architecture of GAN model

decided for the initial run.

Dataset used for classification model is 1413 audio files of real data and fake data. These are 2-second audios sample which is like an audio message on a phone. The dataset is taken as a zip file and then further extracted as a code. The data is loaded and pre-processed.

C. Data Preparation

In the next step, Dataset is split into training and validation in a ratio of 80:20 which translates to 6400 records for training and 1600 records for validation. Data is check for any inconsistency and it is padded and trimmed to meet the required length. After this data is normalized for faster training. After this data is converted into array and the so it can be provided as an input to the models.

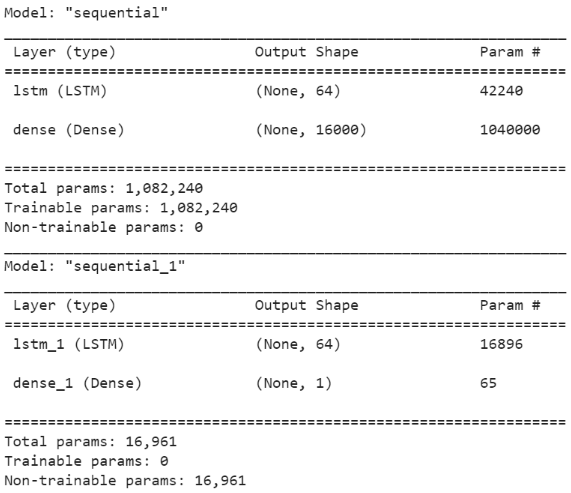
After the classification data is loaded and pre-processed, first spectrogram is generated and visualized of the audio file to listen to the audio file and know the waveform shape. Then data is pre-processed and splitted into test and training datasets. Going forward CNN model is implemented then the model is compiled and trained and evaluated using approriate metrics.

D. Model Architecture

The architecture of GAN model in described in the figure 2. It is built on two neural network, one is generator which takes latent random variable as input and generate random noise. Another is a discriminator which is trained on real audio samples and tries to differentiate between audio samples generated by generator as real or fake. Every time the discriminator successfully detects the fake audio correctly, the generator gets feedback and updates its code to perform better. If the discriminator fails to identify the fake that and categorizes a fake sample as real, the discriminator updates its code.

The different layers in generator as follows:

1) **Input Layer**: The input layer takes up npise vector as input which are the random arrays. This acts as a starting point for the generator to start.
2) **LSTM Layer**: As the audio data is sequential in nature the next layer in LSTM that captures sequential data effectively.It helps in learning the patters and intervals for audio data generation.
3) **Dense Layer**:The final output layer is The dense layer through which the LSTM output is passed. As this layer have linear activation function it produces the final output of generator.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 64)                42240

 dense (Dense)               (None, 16000)             1040000

=================================================================
Total params: 1,082,240
Trainable params: 1,082,240
Non-trainable params: 0
_________________________________________________________________
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_1 (LSTM)               (None, 64)                16896

 dense_1 (Dense)             (None, 1)                 65

=================================================================
Total params: 16,961
Trainable params: 0
Non-trainable params: 16,961
```

Fig. 3. Model Summary for Generator and Discriminator

The layer for discriminators are as follows:

1) **Input Layer**: The discriminator is trained on real data and the array of audio is provided as input to the input layers
2) **LSTM Layer**: As discriminator is being trained on audio data, LSTM layer is a suitable choice to handle the sequential data and adept the learning features.
3) **Dense Layer with Sigmoid Activation**: The LSTM output is fed into a dense (fully connected) layer with a sigmoid activation function. The sigmoid function label the output values between 0 and 1. This is a crucial step where the output generated by generator is classified as real or fake by discriminator..

Figure 3 shows the model summary for Generator and Discriminator. After building generator and discriminator, GAN model is build by stacking them. Using this a pipeline is created in which generator output acts as an input for discriminator though discriminator is set to untrainable in order to ensure faster training of generator and to avoid discriminator becoming very good at identifying real vs fake very quickly[19]. This setup forms the basis of adversarial training.

On the other hand, classifier model is built to differentiate between real and fake audio. CNN model is used for the audio classification purpose.It can automatically learn from spectrogram instead of manual extraction. It is a significant model as it captures local and global patterns. Usually, CNN is used for image classification, however, in this study, spectral visuals are taken as images to classify the model.

Figure 4,gives an overview of the architecture of CNN model. Initially, the audio files are passed as input, post that loaded data is processed and converted into spectrogram Then these samples pass through different CNN layers which are made up of kernels. This helps to detect different features of

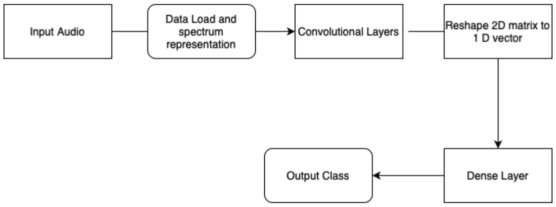

Input Audio
Data Load and spectrum representation
Convolutional Layers
Reshape 2D matrix to 1 D vector
Dense Layer
Output Class

Fig. 4. Overview of CNN model

the waveform. Then clustering layers reduce spatial size and only important features are kept using max pooling. Then flat layer reshapes the 2d matrix to a 1d vector and passes it to the dense layer. Then it is moved to the output layer where the softmax activation function converted the output layer to probability.

E. Training Loop

The training loop is the crucial step in any GAN training. In this research, no pretrained models where used as the experiment is about using LSTM for building GAN, hence the training is carried out from scratch which is computationally exhaustive task based on the resources available. The prepared data is shuffled before feeding into the discriminator for training using train_ds.shuffle(). The real and generated data are concatenated and the labels are assigned as 1 for real and 0 for generated audio. Discriminator weights are updated using discriminator.train_on_batch() to minimize loss. For generator, the label is set to 1 for tricking the discriminator weights are updated using gan.train_on_batch(). This process is repeated till the number of epochs is trained or the condition for early stopping is met.

For classifier model also a training loop is executed to train the model is trained and Adam optimizer is used to optimize the loss.

IV. Evaluation/Results

Models created for this research have been evaluated on various parameters and in this section the results for LSTM-GAN model and the additional classification model will be discussed in detail. Figure 5 displays the graph between discriminator and generator loss. This graph shows that as the training continues the loss of discriminator is increasing and loss for generator is decreasing which means generator is getting better with time. Where as the discriminator is Figure 6 shows the waveform of the generated audio which when compared with Figure 1 clearly shows that more training and analysis on GAN model is needed. It can be observed that in Figure 1, the audio waveform is concentrated between the sample index of 8000 to 12000 while figure 6 is the audio noise generated audio from the implemented GAN model, the amplitude value range is from -10000 to 10000. Both the audio duration is of one second presenting the core of this study. The generated audios are saved in a root folder named 'Generated audio' and can also be played back using the play button after each epoch during training as depicted in figure 7 below.

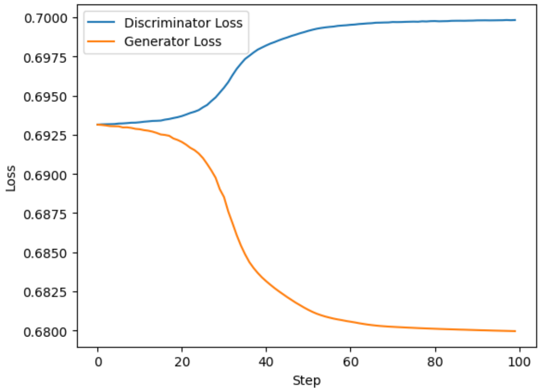


Fig. 5. Discriminator and Generator loss

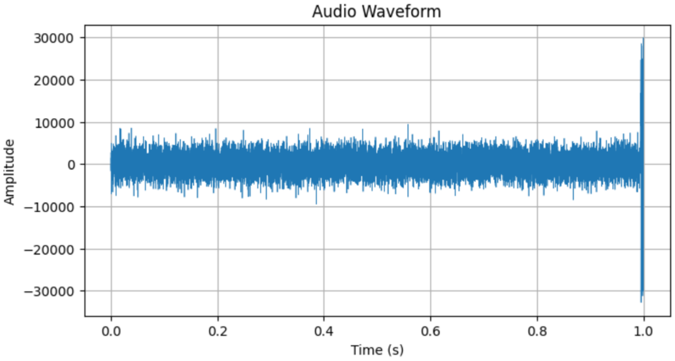


Fig. 6. Waveform for Generated Audio

Through the code we applied a simple feature by including a play button. This button allows us to listen to both the audio that the input audio and generated audio. By simply pressing the button we could listen to the unique features of the input feature and generated audio of the generated Furthermore, the classification model that distinguishes between real and fake audio samples is evaluated based on accuracy, recall, precision, and F1 score. In this audio detection model to evaluate if the model is working as expected few evaluation methods has been applied.

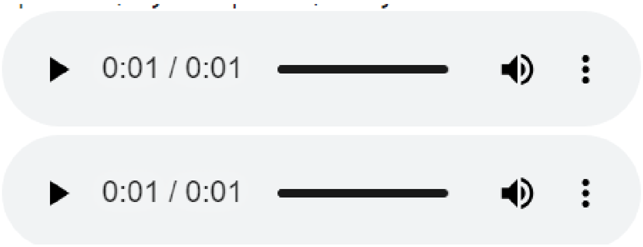


Fig. 7. Playback button for playing audio

A. Model Accuracy

To find the accuracy of the model the accuracy has been compared to the predicted cases over the total number of cases. In this the value of Test Loss: 0.2276 Test Accuracy is 0.9619.

However, we cannot be completely dependent on accuracy as it is not always correct. These two graphs are showing the graph of loss and accuracy.

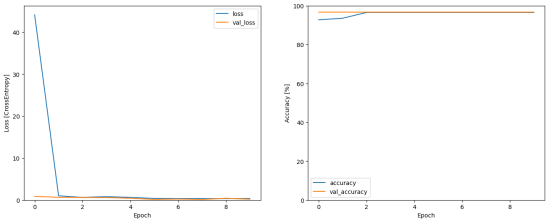


Fig. 8. Accuracy and Loss Graph

B. Confusion Matrix

The next method used is the confusion matrix. In this predicted classes and actual class is present along with their counts.

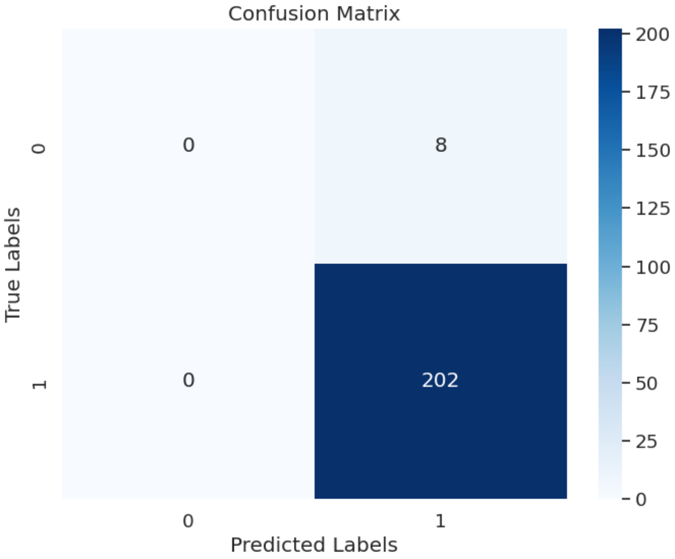


Fig. 9. Confusion Matrix

C. Precision and Recall

The precision model is used for predicting positive values. In this case, high precision is coming which is approx. .96 and Recall is 1.0 which means high chances to find real audio.

D. ROC and Area Under ROC Curve

The model is evaluated using one of the statistical parameters, that is AUC-ROC curve. The area of ROC curve is 0.61 as shown in figure. The value 0.61 indicates that the model can distinguish between positive and negative sound cases somewhat. In other words, the model has biased power but not much stronger. Figure 10 shows the ROC curve the true positive rate is on y-axis and a false positive rate on the x-axis.

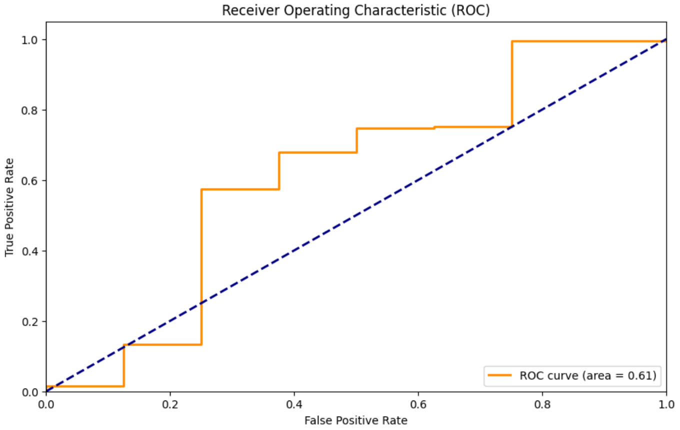


Fig. 10. ROC Curve

V. Conclusions and Future Work

In this research, a GAN model was developed using LSTM layer and it was observed that is in computationally really demanding to train this model. After training on more than 6000 samples also the results generated by the model are white noise only. Although the model architecture can be further enhanced by introducing more layers and training more number of epochs. This process will require more time and resources to complete but this research shows that LSTM can be used in building GAN models for audio generation. On the other hand, the classifier model build using CNN is showing promising results in identifying real and fake audio. The model is moderate but not highly accurate in classifying positive cases as compared to negative cases .This can further be enhanced by training the model on diverse dataset of different lengths like news speech, interviews etc. Another interesting area of further research could be to combine both these models once GAN start producing convincing fakes and evaluate the results of classifier by giving generated audio from GAN as input to classifier.

References

[1] S. Mertes, A. Baird, D. Schiller, B. W. Schuller and E. André, ”An Evolutionary-based Generative Approach for Audio Data Augmentation,” 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2020, pp. 1-6, doi: 10.1109/MMSP48831.2020.9287156.

[2] H. K. Vedamurthy, R. V, and G. S P, “A reliable solution to detect deepfakes using Deep Learning,” IEEE Xplore, Dec. 01, 2022. https://ieeexplore.ieee.org/document/10058638 (accessed Aug. 16, 2023).

[3] Z. Almutairi and H. Elgibreen, “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Algorithms, vol. 15, no. 5, p. 155, May 2022, doi: https://doi.org/10.3390/a15050155.

[4] L. DoraM.Ballesteros, Y. Rodríguez-Ortega, D. Renza, and G. Arce, “Deep4SNet: deep learning for fake speech classification,” Expert Syst. Appl., 2021, doi: https://doi.org/10.1016/j.eswa.2021.115465.

[5] C. Donahue, J. Mcauley, and M. Puckette, “ADVERSARIAL AUDIO SYNTHESIS.” Available: https://arxiv.org/pdf/1802.04208v3.pdf

[6] F. Rong, “Audio Classification Method Based on Machine Learning,” 2016 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS), Dec. 2016, doi: https://doi.org/10.1109/icitbs.2016.98.

[7] S. Mertes, A. Baird, D. Schiller, B. W. Schuller, and E. André, “An Evolutionary-based Generative Approach for Audio Data Augmentation,” IEEE Xplore, Sep. 01, 2020. https://ieeexplore.ieee.org/document/9287156 (accessed Aug. 16, 2023).

[8] “Deepfake Video Detection Using Recurrent Neural Networks,” IEEE Conference Publication — IEEE Xplore, Nov. 01, 2018. https://ieeexplore.ieee.org/document/8639163

[9] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, Emotions Don’t Lie. 2020. doi: 10.1145/3394171.3413570.

[10] Bayat, Nicky Khazaie, Vahid Reza Keyes, Andrew Mohsenzadeh, Yalda. (2021). Latent Vector Recovery of Audio GANs with Application in Deepfake Audio Detection. Proceedings of the Canadian Conference on Artificial Intelligence. 10.21428/594757db.1ee3922d.

[11] Z. Almutairi and H. ElGibreen, “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Algorithms, vol. 15, no. 5, p. 155, May 2022, doi: 10.3390/a15050155.

[12] A. Maccagno, A. Mastropietro, U. Mazziotta, M. Scarpiniti, Y.-C. Lee, and A. Uncini, “A CNN Approach for Audio Classification in Construction Sites,” in Springer eBooks, 2020, pp. 371–381.

[13] M. Mcuba, A. Singh, R. A. Ikuesan, and H. Venter, “The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation,” Procedia Computer Science, vol. 219, pp. 211–219, Jan. 2023, doi: 10.1016/j.procs.2023.01.283.

[14] Z. Ba et al., Transferring Audio Deepfake Detection Capability across Languages. 2023. doi: 10.1145/3543507.3583222.

[15] Yi, Jiangyan Tao, Jianhua Fu, Ruibo Yan, Xinrui Wang, Chenglong Wang, Tao Zhang, Chu Zhang, Xiaohui Zhao, Yan Ren, Yong Xu, Le Zhou, Junzuo Gu, Hao Wen, Zhengqi Liang, Shan Lian, Zheng Nie, Shuai Li, Haizhou. (2023). ADD 2023: the Second Audio Deepfake Detection Challenge.

[16] A. Moussa and H. Watanabe, “Audio Translation with Conditional Generative Adversarial Networks,” IEEE Xplore, Feb. 01, 2020. https://ieeexplore.ieee.org/document/9065067 (accessed Aug. 16, 2023).

[17] S.-H. Lee, J.-H. Kim, K.-E. Lee, and S.-W. Lee, “FRE-GAN 2: Fast and Efficient Frequency-Consistent Audio Synthesis,” IEEE Xplore, May 01, 2022. https://ieeexplore.ieee.org/document/9746675 (accessed Aug. 16, 2023).

[18] Narasimha Reddy Yella and B. Rajan, “Data Augmentation using GAN for Sound based COVID 19 Diagnosis,” Sep. 2021, doi: https://doi.org/10.1109/idaacs53288.2021.9660990.

[19] Y. Li, L. Gao, Z. Tang, Q. Yan and Y. Huang, ”A GAN-Based Feature Generator for Table Detection,” 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 2019, pp. 763-768, doi: 10.1109/ICDAR.2019.00127.