Deepfake Audio Detection System

MSc in Data Analytics(MSCDA-B)
Data mining & Machine learning 2 – Project Report

Himanshu Gupta	Sodiq Ololade	Ashish Rawat
x18203302	x19133979	x18185801

School of Computing
National College of Ireland

Supervisor: Prof. Anu Sahni

DeepFake Audio Detection System

Abstract*—The problem of recognizing fake or computer-generated audio requires identifying where it differs from real audio. Audio files contain subtle differences which are imperceptible to the human ear. Also, advancements in the field make of speech synthesis make it impractical to try picking up on differences between these audio classes without first transforming the audio file. In this research, we use a dataset of pre-recorded audio files containing both fake audio and real audio. We transform the audio into spectrograms to gather information on the frequency spectrum of an audio file as it varies with time. Spectrograms are visual representations of audio files which capture intensity and frequency of speech. These are used as input to a Convolutional Neural Network (CNN) to train a model on classifying the data. In this research, CNN uses three hidden layers with a categorical cross-entropy loss function and an 'Adam' optimizer to achieve the results. We obtained an accuracy of (78%). In practice, the effects of an incorrect classification could be far reaching, thus making a high accuracy desirable. Though a high accuracy is achieved, we found the process to be time consuming due to the time taken to convert the audio to spectrograms. The outcome of this research is a spectrogram-based audio classifier.

***Keywords*—CNN, spectrogram, keras, spoofing, ASVSpoof 2019**

I. Introduction

Voice cloning technology has advanced rapidly over the years, finding application in many industries such as security and biometric verification, medicine, entertainment and gaming. With this increased use, there is growing concern about the security challenges which come with the vulnerabilities exposed by such technology [1].
Along with voice cloning, computer-generated audio has advanced, and it has become increasingly difficult to tell real speech from synthesized or computer-generated speech. This opens the world to a new plethora of attacks by fraudsters aiming to manipulate unsuspecting individuals for their objectives. Equally speedy response in the security field is required combatting this problem [2].

Previous works of this nature use extracted features such as the Mel's Frequency Cepstral Coefficients, Constant Q Cepstral Coefficient, Electric Network Frequency among others to generate numerical representations of audio data and then use these data to train a model such as the Support Vector Machine (SVM) classifier and the Hidden Markov Model (HMM) [3][4][5].

Our response to this problem is the use of machine learning techniques to identify real and computer-generated audio, training a model to tell the difference even if the human ear cannot. To achieve this, we make use of Convolutional Neural Network (CNN) classifier model. The CNN is a Deep neural learning technique which mimics the learning process of the human brain to train a model. It works using image data as model input. We begin by converting audio samples in a dataset to spectrograms and then feed this into the CNN model by way of training the model. We then use CNN to classify these files into real and fake audio files. We adjusted the tuning parameters to improve the results of the model. We evaluate the model using the accuracy, recall and precision of the model.

II. Related work

[6] This paper analyzes the spoof audio created by recording the speech and replay it which dodge an automatic speaker verification mechanism. There is not much research conducted on building spoofing detection system which can handle the vulnerability of ASV systems. The detection system is based on the convolutional neural network. Similar ASVspoof dataset considered in this paper but with the 2017 version. The training dataset consists of a speaker, genuine and replays audio clips. State of the art LCNN machine learning method used which consist of five convolutional layers, four network layers, five max pool layers and two fully connected layers. RELU is used as an activation function as it was a state of the art as compared to Max-feature-map (MFM) activations. For data preprocessing mean-variance normalized spectrogram generated. To keep the input consistent audio clips are truncated to 4 seconds before putting into the convolutional neural network. The output classes are spoofing, and genuine so binary entropy used, and it needs to be optimized by training the network. Maximum 100 epochs used while training the network and equal error rate calculated for the overall dataset. Further, they used SLIME model to gain insights about the model and then for the model prediction. SLIME is works based on interpretable sequence for each input to get the class explanation. It was concluded that the performance of spoofing detection system majorly depends on the first few audio samples and this model got the satisfactory results.
One of the biggest challenges of automatic speaker verification system is the replay attack. The vulnerability of the ASV system against such attacks studied with experimentation. Replay attacks can easily be induced with smart devices. This study proposed a model which overcome the problem of earlier models where different features of spoofing and genuine audio clips were not extracted efficiently. The convolutional neural network of 1D layer used for building model and take the ASVSpoof 2017 challenge dataset for the same. Instead of using spectrograms raw waveforms of audio files at 16 kHz used for this. The standard length of audio files considered before putting it into CNN. For evaluating the performance equal error rate calculated which were different for different frequencies [7]. On the other hand, some studies used traditionally and auto encoded generated robust audio features. Dataset is ASVSpoof 2017 challenge and focus was on ensuring robustness by finding the relevant features based on Gaussian mixture model-universal background model. The error rate calculated on the known audio features and machine-learned features by using the fused model. The hybrid model performed very well as compare to individual ones [8]. Not only the automatic speaker system gets affected by spoofing attacks via replayed or synthesized data. Some acoustic

speaker recognition system also faced the same problems. This problem studied and researched to solve by using audio-visual speaker recognition. Data synchronicity is verified simultaneously with the help of proposed coupled hidden Markov models. The detection system of spoofing build is text-dependent which provide new information about transcriptions with the help of new features. These features get evaluated by separately created dataset consist of 30 speakers and each speaker contains 270 utterances and different spoofing scenarios were considered [9].

[10] this study used hybrid feature extraction method based on the segments. Feature extraction consists of two parts one is Mel-frequency cepstral coefficient which is short term power spectrum of audio and second Constant-Q cepstral coefficients (CQCC) features which are commonly used in spoof speech detection system. Three Deep learning models are used to train the network which is LSTM, DenseNet and the hybrid of DenseNet-LSTM architectures. The hybrid model with mixed features outperforms the others. The model trained using train subset and evaluate using dev subset. The hybrid model based on different segment consist of different background audio of different devices make this best suitable model for the spoofing detection and on the other hand, some studies considered spoofing countermeasures based on the average inter-frame difference which is used to distinguish between genuine and synthetic speech. The effectiveness of this method depends on the global variance and dynamic variation of speeches. Voice cloned SAS database is used which contains voices of native British speakers with the inclusion of non-parallel utterances. Gaussian mixture model, joint factor analyses and PLDA are used for speaker verification system and all three are vulnerable to all the spoofing techniques[11].

[12] Biometric security systems also depend on speech sentences and visual inputs were to identify the authorized speech is a big challenge. Deep recurrent neural networks applied which consist of multiple recurrent network layers to achieve the results. ASVspoof 2015 dataset is used for this study. The model performs expectedly with promising error rates. With the similar dataset in one of the study, a joint model proposed, convolutional LSTM and CLDNN combined to build the spoofing detection which outperforms the previous study on BTAS 2016 dataset and raw waveforms input considered which gave better results [13]. Gated recurrent convolutional network model also used to achieve the same. It is a deep feature extractor which shows signals as utterance-level embeddings. Signal to noise technique used for feature extraction. The evaluation performed on logical access attacks on noisy audio and clean audio to detect replay attacks [14]. The performance of these models optimizes in the clean environment but it degrades if the environment is noisy and not up to the mark for training the model and this problem overcomes by using deep feedforward neural network and feed the input with all distinguish and robust features. Multiple types of training proposed like dropout training to handle the noise and avoid overfitting. Results showed that performance has increased and the error rate decreased from 19% to 2%with these parameters [15]. One of the earlier papers proposed research in the area of electronic signal for building an anti-spoofing system such as digital speech watermark which is based on the encoded and decoded concept. Watermark was infiltrated in the signal at one end and then extracted on the other hand to identify if it is authorized speech and results showed that it was successful in case of anti-spoofing attacks systems [16]. To avoid the overfitting of the traditional convolutional model and low performance of the Gaussian model this study used DenseNet architecture to obtained the best results. This work also is done on ASVspoof dataset and the used MFCC and CQCC techniques to distinguish between spoof and genuine audio features as shown in figure 1 provided in this research paper. Overall the hybrid model performed 46% times better than the baseline system [17].

[18] Experimental study on the ASVspoof 2017 data for building the anti-spoofing system. Long term average spectrum used and extracted features from then using deep neural network classifier and comparative study performed with the gaussian mixture model with MFCC used above and DNN classifier performed much better than baseline Gaussian model because long term average spectrum served better features as compare to MFCC based features. One mode researched attempted using the gated recurrent neural network in combination with long short term memory and outperforms Gaussian model with the equal error rate of 9.81% ASVspoof 2017 data and it also concluded that recurrent network worked better with sequential model unlike Deep neural feed-forward model [19].

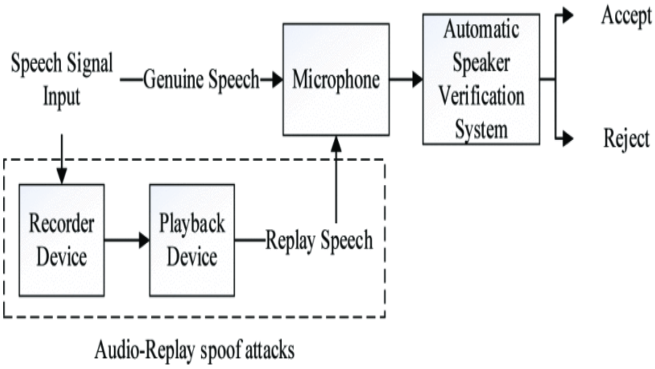

Figure 1 Audio-Replay attacks in ASV system

III. Research methodology

For data mining, two kinds of methodologies are there one is KDD (Knowledge Discovery and Data Mining) and one is CRISP (Cross-industry standard process). For carrying out this research KDD is followed as shown in figure 2.

With the advancement of science and technology data becomes a very crucial part of any industry. The source of data is different and available in many formats. It is very important to deduce the meaningful information from the data which helps in making any product more efficient and better. In this process, the raw data is pre-processed first according to the specific needs and then after transformation, it passes to the data mining for analyzing any existing patterns and relationships among the data variables. Finally, after looking at the patterns various interpretations occur. This section explains how KDD achieved in this research.

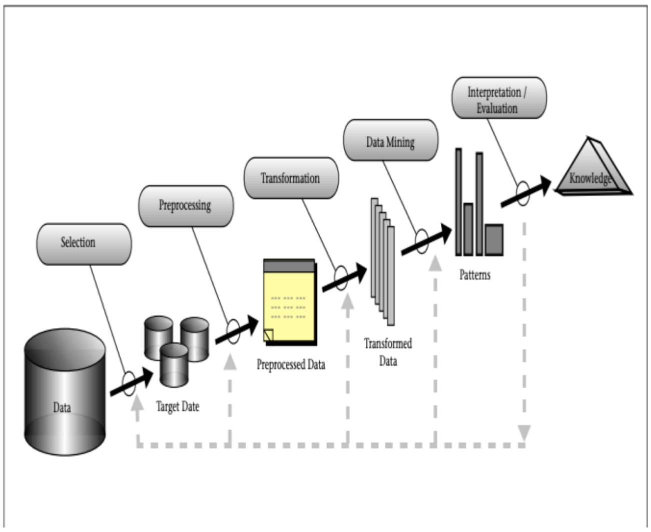

Figure 2 KDD Process Flow

A. *Data Gathering and Preprocessing*

For building the spoofing detection system many earlier researches used ASVspoof challenge dataset of different versions. In this study Automatic speaker verification ASVspoof 2019 dataset used which is publicly available. It consists of several audio files which contains real, text to speech and voice conversion audio data [21]. The data is downloaded and stored in the hard disk directory and referred from there. Due to computational constraints of the system, limited audio files (25380 audio files) have been used in this research along with the given text file containing description of each audio files. We utilized the text files to segregate the audio files according to their system id (A01- A19) and created three folders, one for each class (Real, Spoof_TTS and Spoof_VC). Audio files have not been standardized in the processing part to maintain the essential feature of the data as clips have only few seconds audio. For each audio class a respective image folder is created which contains the spectrograms of audio files belonging to that class. And finally, all the spectrograms are moved to a single directory to create dataframe of image data.

B. *Applied Methods*

After the successful import of all the libraries the generated spectrogram images are feed as an input in the machine learning model. This is supervised multiclass classification problem and the output dependent classes are Real, Spoof_TTS (text to speech spoof) and Spoof_VC (voice conversion).

The deep learning model is an extension and created over the foundation of the convolutional neural network. In this research proposed model built based on the Keras framework. Keras is deep learning library which is purely developed in the python programming language and it can on the top of TensorFlow or Theano. TensorFlow is an API based on the AI system provided by Google and it has multiple dimensions which go from one end of the data flow chart to the other end. It works as a backend for Keras framework [22]. The general form of CNN is shown in figure 3.

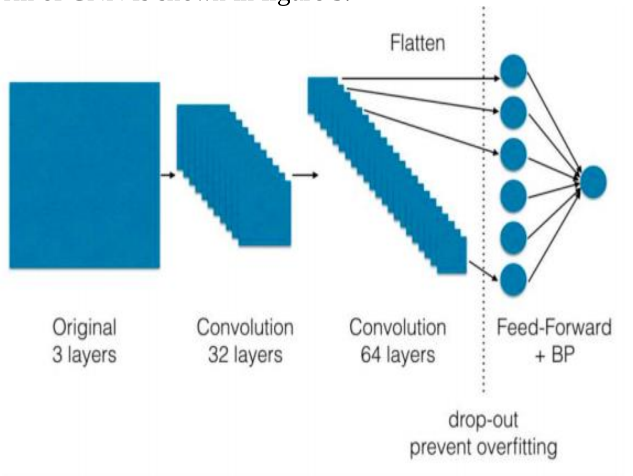

Figure 3 General CNN model

In this study a sequential model created with three Conv2D consisting of 32,64 and 128 layers. Batch normalization layers also added which convert the input values to standardized values automatically in Deep neural network. After normalization, max pooling layer added in the network which reduces the dimensionality of the input images features while keeping the maximum spatial data. The layer models look like the below figure 4

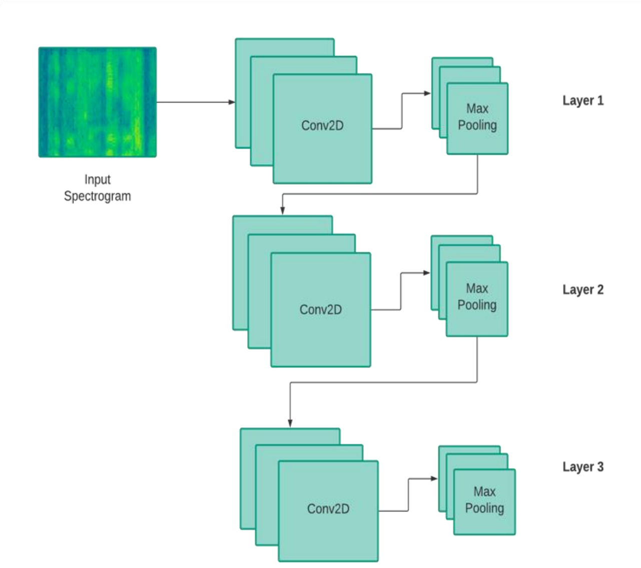

Figure 4 Keras CNN model layers

ReLU is used as an activation function in conv2d layers and it gave the nonlinear output. and model compiled with ‘categorical_crossentropy’ because output classes are categorical and have more than 2 labels that are why binary cross-entropy is not used here. The goal here is to keep the loss minimum to get the optimized results, the loss is a cost which explained the values of variables associated with real integers lost. Adam is such loss function used here to gain the optimization. The model summary tells how many total overall parameters got trained in the model. Data is split into train and test data in 80:20 ratio using train_test_split of sklearn model selection. After splitting the training data again split into train and validation dataset in 90:10 ratio which used as validation while training the model. Keras provides image processing classes for the image data augmentation used in the model for image data generator class. This class created for train, validation and test data.

The model got trained with 50 epoch and due to the high number of epoch there is a possibility of overfitting of the model and that is why Early stopping used which stops the training if the model is not improving further on a holdout validation set. It found for some cases it stopped at epoch 21. Learning rate also used which determined how fast the model learning the problem. Both the learning rate and Early stopping passed as callbacks in the model. After model training has completed test data is used for model prediction and visualize that how much data points accurately predicted in the respective three output categories.

C. *Evaluation Methods*

For evaluating the model accuracy and the degree of fit several methods used such as validation loss, validation accuracy and training accuracy. During the model training with the increasing number of epochs, the validation loss should be decreasing and the accuracy should be increasing. If the validation accuracy gets starts decreasing and loss increasing then it means that model is not learning it just cramming. On the other hand, in the case of overfitting, both might be increasing.

The ideal situation for model fitting is that the training loss and validation loss should be equal. If training loss is far higher than validation loss then it means the model is underfitting and if the training loss is lesser than validation loss then the model is overfitting. The aim is to achieve ideal results. Results of validation loss and accuracy explained in the result section.

D. *System specifications*

Table 1: Hardware Specification

Processor	Memory	Speed
Intel Core i5 / GPU	8 GB	1.8Ghz

Table 2: Software Specification

Storage	Software and Libraries
Local hard disk or Cloud Storage of minimum 2.2 GB	Python, pandas, keras, soundfile, numpy, scipy, random, pathlib, shutil, random, sklearn

IV. Evaluation and Results

From the experiments carried out, we recorded a final training accuracy of 81.47% and training loss of 46.19% at 21 epochs, beyond which the model did not improve any further. Concurrently, we recorded a validation loss of 59.75% and an accuracy of 77.98%. Below is a plot of the progression of both the loss and the accuracy per epoch for both validation and training.

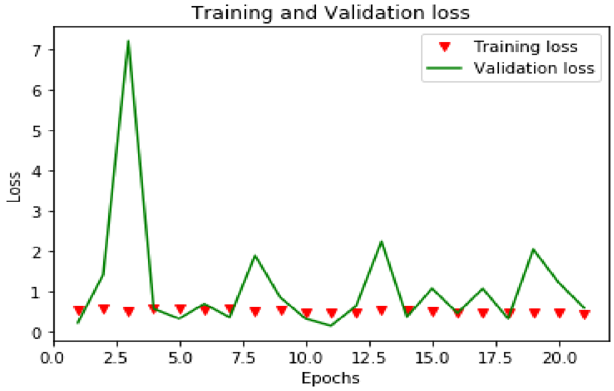

Figure 5 Training and Validation Loss per Epoch

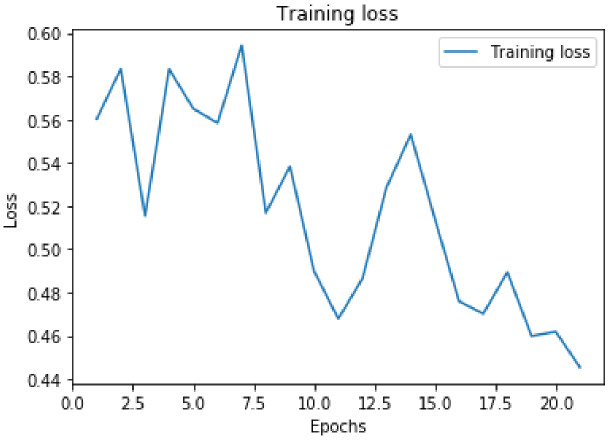

Figure 6 Training Loss per Epoch

In figure 5, we see a graph showing the training validation loss for each epoch. This is the actual accuracy we need from our results. It shows that when we run the model through a set of images for validation, we achieve an accuracy of 77.98%. This shows there is much room for improvement of the model, with our earlier projections for accuracy requiring the above 90% accuracy. We also see a training accuracy of 81.47% which shows how well the model fits to the training data.

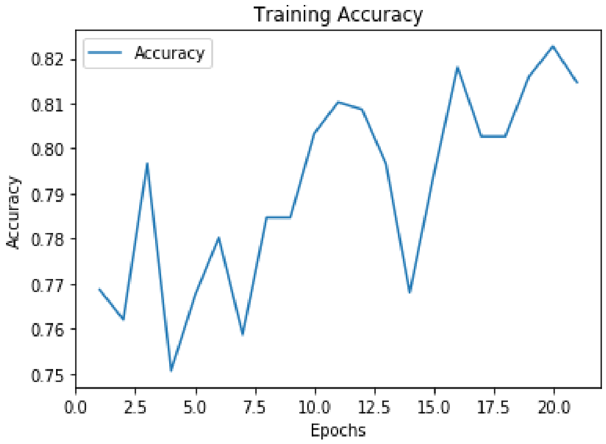

Figure 7 Training Accuracy per Epoch

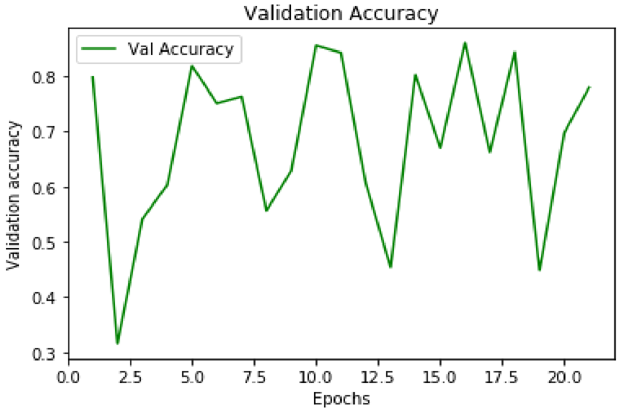

Figure 8 Validation Loss per Epoch

For the test data we achieved the prediction accuracy of 85%(approx.) . It can be observed from comparing figure 9 and 10 that the model performed relatively better in classifying spoof data from Real data . The low performance in classifying the real class is observed due to availability of less data of real class.

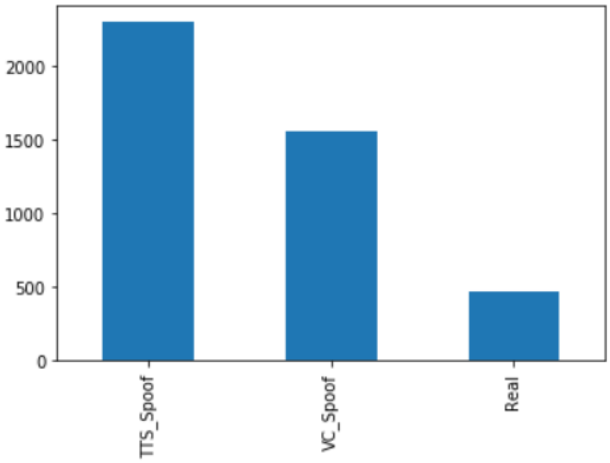

Figure 9 Original Label of Test Data

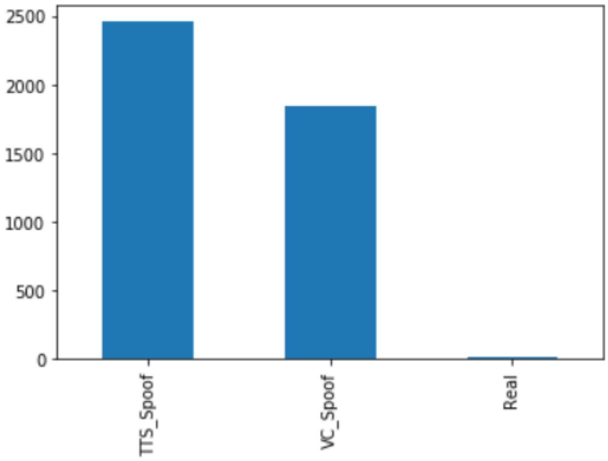

Figure 10 Predicted Label of Test Data

V. Conclusion and Future Work

In this paper we explored and implemented Keras CNN based model for detecting fake audio using the ASVSpoof 2019 database. We have mainly considered noiseless audio files for training purpose. However, due to the nature of the vulnerabilities of systems to deep fake attacks, continuous improvement on the response side is required and lower accuracy cannot ensure confidence in this method in practical systems.

Also, real life scenarios sometimes require almost real time detection of deep fakes. The time taken to convert from audio to a spectrogram may prove a hurdle, limiting the practical use of this approach. To mitigate this, future work may consider exploring methods to speed up the conversion of audio files to spectrograms. We may also consider the use of ensemble methods to improve the accuracy of the model. We will also consider the use of audio signals which contain noise to train the model and to improve the practical applicability of the model in real world scenarios.

REFERENCES

[1] M. Dordevic, M. Milivojevic and A. Gavrovska, "DeepFake Video Analysis using SIFT Features", 2019 27th Telecommunications Forum (TELFOR), 2019. Available: 10.1109/telfor48224.2019.8971206 [Accessed 26 April 2020

[2] P. Korshunov and S. Marcel, "Vulnerability assessment and detection of Deepfake videos", 2019 International Conference on Biometrics (ICB), 2019. Available: 10.1109/icb45273.2019.8987375 [Accessed 26 April 2020].

[3] S. Jadhav, R. Patole and P. Rege, "Audio Splicing Detection using Convolutional Neural Network", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019. Available: 10.1109/icccnt45670.2019.8944345 [Accessed 26 April 2020]

[4] Z. Ali, M. Imran and M. Alsulaiman, "An Automatic Digital Audio Authentication/Forensics System", IEEE Access, vol. 5, pp. 2994-3007, 2017. Available: 10.1109/access.2017.2672681 [Accessed 26 April 2020].

[5] T. Bhangale and R. Patole, "Tampering Detection in Digital Audio Recording Based on Statistical Reverberation Features", Advances in Intelligent Systems and Computing, pp. 583-591, 2019. Available: 10.1007/978-981-13-3600-3_55 [Accessed 26 April 2020].

[6] B. Chettri, S. Mishra, B. Sturm and E. Benetos, "Analysing The Predictions Of a CNN-Based Replay Spoofing Detection System", *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018. Available: 10.1109/slt.2018.8639666 [Accessed 26 April 2020].

[7] S. Shukla, J. Prakash and R. Guntur, "Replay attack detection with raw audio waves and deep learning framework", *2019 International Conference on Data Science and Engineering (ICDSE)*, 2019. Available: 10.1109/icdse47409.2019.8971793 [Accessed 26 April 2020].

[8] B. Balamurali, K. Lin, S. Lui, J. Chen and D. Herremans, "Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features", *IEEE Access*, vol. 7, pp. 84229-84241, 2019. Available: 10.1109/access.2019.2923806 [Accessed 27 April 2020].

[9] L. Schonherr, S. Zeiler and D. Kolossa, "Spoofing detection via simultaneous verification of audio-visual synchronicity and transcription", *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017. Available: 10.1109/asru.2017.8268990 [Accessed 27 April 2020].

[10] L. Huang and C. Pun, "Audio Replay Spoof Attack Detection Using Segment-based Hybrid Feature and DenseNet-LSTM Network", *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. Available: 10.1109/icassp.2019.8682573 [Accessed 27 April 2020].

[11] Z. Wu et al., "Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768-783, 2016. Available: 10.1109/taslp.2016.2526653 [Accessed 27 April 2020].

[12] S. Scardapane, L. Stoffl, F. Rohrbein and A. Uncini, "On the use of deep recurrent neural networks for detecting audio spoofing attacks", *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017. Available: 10.1109/ijcnn.2017.7966294 [Accessed 27 April 2020].

[13] H. Dinkel, Y. Qian and K. Yu, "Investigating Raw Wave Deep Neural Networks for End-to-End Speaker Spoofing Detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2002-2014, 2018. Available: 10.1109/taslp.2018.2851155 [Accessed 27 April 2020].

[14] A. Gomez-Alanis, A. Peinado, J. Gonzalez and A. Gomez, "A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 1985-1999, 2019. Available: 10.1109/taslp.2019.2937413 [Accessed 28 April 2020].

[15] Y. Qian, N. Chen, H. Dinkel and Z. Wu, "Deep Feature Engineering for Noise Robust Spoofing Detection", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1942-1955, 2017. Available: 10.1109/taslp.2017.2732162 [Accessed 28 April 2020].

[16] M. Nematollahi, S. Al-Haddad, S. Doraisamy and M. Ranjbari, "Digital speech watermarking for anti-spoofing attack in speaker recognition", *2014 IEEE REGION 10 SYMPOSIUM*, 2014. Available: 10.1109/tenconspring.2014.6863080 [Accessed 28 April 2020].

[17] L. Huang, Y. Gan and H. Ye, "Audio-replay Attacks Spoofing Detection for Automatic Speaker Verification System", *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2019. Available: 10.1109/icaica.2019.8873465 [Accessed 28 April 2020].

[18] B. Bakar and C. Hanilci, "An Experimental Study on Audio Replay Attack Detection Using Deep Neural Networks", *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018. Available: 10.1109/slt.2018.8639511 [Accessed 28 April 2020].

[19] Z. Chen, W. Zhang, Z. Xie, X. Xu and D. Chen, "Recurrent Neural Networks for Automatic Replay Spoofing Attack Detection", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. Available: 10.1109/icassp.2018.8462644 [Accessed 28 April 2020].

[20] U. Fayyad, Advances in knowledge discovery and data mining. Menlo Park: AAAI Press, 1996.

[21] J. Yamagishi et al., "ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database", Datashare.is.ed.ac.uk, 2019. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/3336. [Accessed: 22- Feb-2020].

[22] A. LI, Y. LI and X. LI, "TensorFlow and Keras-based Convolutional Neural Network in CAT Image Recognition", *DEStech Transactions on Computer Science and Engineering*, no., 2017. Available: 10.12783/dtcse/cmsam2017/16428.