IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.

EDL-Det: A robust TTS detector using VGG19-based YAMNet and Ensemble Learning Block

Rabbia Mahum[1], Aun Irtaza[1], and Ali Javed[3]

[1,2] Computer Science Department, UET Taxila, 47050, Pakistan.

[3] Software Engineering Department, UET Taxila, 47050, Pakistan.

Corresponding author: Rabbia Mahum (e-mail: rabbia.mahum@uettaxila.edu.pk).

This research work was supported by University of Engineering and Technology Taxila, Pakistna.

ABSTRACT Various algorithms exist for the audio deep fake synthesis, such as deep voice, tacotron, fastspeech, and imitation techniques. Despite the existence of various spoofing speech detectors, they are not ready to distinguish unseen audio samples with high precision. In this study, we suggest a robust model, namely Ensemble Deep Learning based Detector (EDL-Det) to detect text-to-speech (TTS) and categorize it into spoofed and bonafide classes. Our proposed model is an improved method based on YAMNet employing VGG19 as a base network instead of MobileNet combined with two other deep learning(DL) methods. Our proposed system effectively analyzes the mel-spectrograms generated from input audio to extract the better artifacts underlying the audio signals. We have added an ensemble learning block that consists of ResNet50, and InceptionNetv2. First, we convert speech into mel-spectrograms that consist of time-frequency representations. Second, we train our model using the ASVspoof-2019 dataset. In the end, we classified the audios converting them into mel-spectrograms using our trained binary classifier along with a majority voting scheme by three networks. Due to deep convolutional network architecture, our proposed model effectively extracts the most representative features from the mel-spectrograms. Furthermore, we have performed extensive experiments to assess the performance of the suggested model using the ASVspoof 2019 corpus. Additionally, our proposed model is robust enough to identify the unseen spoofed audios and accurately classify the attacks based on cloning algorithms.

INDEX TERMS Deep learning; DeepFake Audios; Fake Speech; Text-To-Speech Detection; VGG19; Mel-Spectrograms;

I. I.INTRODUCTION

With the advancement in the domain of artificial intelligence, various automatic speaker verification (ASV) systems have been introduced to authenticate users in various applications such as banking, forensic laboratories, call centers, etc. Thus, speech is commonly used as a transmitting medium in digital devices, for example, mobile phones and computers. However, with the advancement of machine learning and deep learning models, it has become very easy to manipulate the signals and generate spoofed speech to deceive the listener [1]. Moreover, various speech synthesis algorithms, such as GAN [2], Deepvoice [3], tacotron2 [4], and wavenet [5], have gained importance to generate natural speech just like humans and defeat the automatic speaker verification (ASV) systems. For example, false information related to politics based on deep fakes became a significant threat to the US presidential election in 2020 [6]. Furthermore, an incident of loss of USD 243,000 has occurred employing an audio deep fake [7] in bank transactions. Therefore, these incidents show the vulnerability of the ASV systems that are used widely in various security systems.

There exist three types of modalities replay attack (RA), text-to-speech synthesis (TTS), and voice cloning (VC). TTS and VC comprise regenerated content and are more similar to natural speech than RA. In ASVspoof 2019 competition, logical access (LA) and physical access (PA) tasks for synthesized speech and RA detection were introduced for developing ASV systems, respectively. Various researchers have proposed different approaches [8] for spoofing detection. Some algorithms exist based on machine learning techniques to discern the audios based on data-driven and knowledge-focused countermeasures [9]. However, in traditional machine learning algorithms, hand-crafted features extraction is performed which is time-consuming task, moreover, they may ignore the deep features underlying the audios spectrograms [10]. With the improvement in the domain of convolutional neural networks (CNNs), some methods have been proposed based on deep layers. For example, [11] developed an end-to-end algorithm employing raw waveforms as input. Moreover, a lightweight convolutional neural network has been employed by [12], namely LCNN utilizing softmax loss function to detect anti-spoofed attacks.

Various fake speech detecting systems have been tested along with ResNet [13] and explored with other classifiers for better performance [14]. The main challenge in existing systems is that they may fail on unseen audios and are less generalized to identify all types of synthesized speeches effectively. Moreover, mostly the existing solutions rely only on accuracy for evaluation that may not provide an authentic picture regarding performance.

Therefore, in this research we propose a novel and robust framework to detect spoofed voices specifically TTS-based using the Customized YAMNet deep learning model alongwith ensemble learning block. The convolutional layers extract the most elusive features from the mel-spectrograms (2-D images) based on VGG19 as a base network in YAMNet than the raw inputs of audios, which is the foundation for precisely detecting spoofed voices.

Our proposed model is mainly divided into three phases. First, audio features have been extracted in the form of images known as mel-spectrograms. Second, a deep layered network has been trained using the ASVspoof-2019 dataset to classify the audio input as fake or real. Third, the network performs the binary classification of mel-spectrograms. To evaluate EDL-Det's performance and effectiveness, we perform our experiments utilizing a publically available dataset, ASVspoof 2019 and ASVspoof 2021. We assessed the performance of the suggested system using PA (replay and bonafide samples) and LA (voice conversion, speech synthesis, and bonafide) sets from ASVspoof 2019 corpus using several standard metrics than accuracy. The major offerings of the proposed system are presented below:

- We propose an improved deep learning model (EDL-Det) based on YAMNet architecture for spoofed audio detection similar to image classification models.
- EDL-Det utilizes VGG19 as a base network in YAMNet to extract the features from mel-spectrograms. Moreover, we attached an ensemble learning block with the main network comprising ResNet50 and InceptionNetV2 to strengthen the final classification decision.
- EDL-Det is a robust speech spoofing detector that can detect several types of spoofing attacks i.e., replay attacks and voice conversion.
- We evaluated our proposed system by employing extensive experiments that confirm EDL-Det's significance over existing techniques.
- We used ASVspoof 2019 dataset for training and evaluation. Moreover, we also cross-validated the performance using deep fake speeches from ASVspoof 2021 dataset. The results show the efficacy of the proposed spoofing detector specifically TTS synthesis.

The remaining paper is ordered as follows: Section 2 defines the related work, Section 3 enlightens the methodology of the proposed technique, Section 4 defines the experiments performed, and Section 5 demonstrates the conclusion and limitations.

III.RELATED WORK

Various models have been proposed to classify audio based on audio features [15, 16]. The applications that can protect ASV systems from attacks are called deepfake speech detectors. Thus, various machine learning and deep learning-based works have been proposed to detect forged speech. In [17], an SVM-based classifier has been utilized as AVS employing GMM. They attained an equal error rate of 4.92% and 7.78% on the 2006 NIST for speaker identification core test. The authors have proposed the Gaussian Mixture Model (GMM), and a Relative Phase shift with a Support Vector Machine (SVM) for the synthetic speech detection to minimize the weaknesses of speaker verification systems. Moreover, a detailed comparison of the Hidden Markov Model (HMM), and DNN has been performed to detect spoofed speech [18]. In[19], the proposed model employs the spectrograms in image form as input to CNN, thus forming a base of audio processing using images. In [20], various features descriptors have been used, such as Mel Frequency Cepstral Coefficient (MFCC), spectrogram, etc., and the effect of GMM-UBM on the accuracy has been analyzed. It is concluded that the combination of different feature descriptors gives better results in terms of Equal Error Rate (EER). Moreover, in the last two decades, text-to-speech systems have become so powerful that they can generate a realistic voice after training limited audio samples from target speakers[21]. Therefore, it is a huge threat to ASV systems as they may be attacked by the naturalness of the speech generated [22].

Moreover, to decline the computational cost of the polynomial kernel SVM by exchanging the dot product among two utterances with two i-vectors [23]. Furthermore, the authors applied the features selection technique, attaining a 64% dimensionality reduction in features with an equal error rate of 1.7% [23]. In comparison, Loughran et al. [24] overcame the issue of imbalanced data (where the one class samples are greater than the other) by utilizing a Genetic algorithm (GA) with an adjusted cost function. Malik et al. [25] developed a system for audio forgery detection based on the environment's acoustic signatures by investigating the audio's integrity. However, these proposed models failed to address synthesized audio content comprehensively. In [26], the bispectral method for analyzing and detecting synthetic voices has been

proposed. They examined uncommon spectral features in fake speeches synthesized using DNNs, which they called bispectral features. They also tried to find high-order polyspectral features to discern the fake audio.
[27] explained that the spectral features are significant for detecting synthetic speech, such as MFCC features are better than other spectral features for the model's input. Furthermore, [28] described the challenges and limitations of the spoofed detection models.

A DNN-based classifier has been proposed to detect and employed highlight Human Log Likelihoods (HLL) as a metric for scoring and proved to be better than classical log-likelihood ratios (LLR) [29]. Additionally, they also utilized various cepstral coefficients for the classifier's training. [30, 31] also employed a convolutional neural network for audio classification. An extensive comparison has been made using DL techniques for fake audio detection in [32], demonstrating that CNN and Recurrent Neural Network (RNN) based models give better results than all other employed techniques. A capsule network-based approach has been proposed by [33]. They enhanced the generalization of the proposed system and examined the artifacts deeply to increase the overall performance of the model. They also investigated the replay attacks in audios employing their network. In [34], authors proposed a model for fake audio detection named DeepSonar. They analyzed the network layers and the activation patterns for various input audios to examine the difference between fake and real speeches. They employed three datasets consisting of English and Chinese language and attained average accuracy of 98.1%.

In [35], the authors have proposed a model for fake audio detection based on micro-features such as voicing onset timing (VOT) and articulation. They analyzed that VOT numbers are high in fake speeches and attained a 23.5% error rate employing a fusion of both features descriptors. The authors claimed that these micro-features could be used as standalone features for fake audio detection. Moreover, Temporal Convolutional Networks (TCN) [36] have outperformed traditional algorithms such as RNNs and LSTMs for various tasks. The latest deep learning techniques for text-to-speech synthesis systems, such as [37], cloned the voice using original speech recordings. It requires a few minutes of recording in real voice and generates fake audio in some seconds. Although the techniques have been improved in [38], they still face the challenge of naturalness. Moreover, VOCo and Double voice models are developed to generate fake speeches based on signal processing mechanisms rather than machine learning algorithms [39, 40]. The generated voice copies the accent, pitch, rhythm, genre, and plain text from the original speech based on mthe apping. The number of fake speeches was dependent on the duration of the real speech. Due to the high similarity between real and fake voices, it becomes very easy to betray the listener, and fake speech could be utilized as evidence in legal matters. Although various models have been proposed to classify audio based on machine learning algorithms, extensive steps are still required, such as audio data pre-processing, hand-crafted features extraction, features selection, and classification that might surge the computation cost and increase human efforts. On the other side, in very few studies CNN has been employed for features extraction and then a traditional classifier is used for classification [41]. Moreover, these existing studies have failed to fully discern the fake voices and thorough evaluation is not performed to evaluate its rtheirstness employing the various manipulated voices (changing pitch, rhythm, anresamplingle it without changing the linguistics). Furthermore, audio artifacts are more difficult to detect than image artifacts as they are easily visible to the eyes. The voice signals are in 1-Dimensional form. Therefore it is not easy to extract the features and utilize them similarly to image artifacts with various channels based on 2-Dimensional spatiality. In addition to this, as indoor or outdoor voices have environment's noises, therefore it is very easy for the fake voice generators to add real-world noise in to voice to make fool the listener or ASV system. Thus, an automatic fake audio detector that is robust enough to identify the fake audio of various environments is still needed.

III.METHODOLOGY

Deep learning architectures are made of various layers, such as input, hidden, and classification layers, as shown in Figure 1. These hidden layers have various types: convolutional, batch normalization, pooling, activation, etc. The convolutional neural networks-based models extract features utilizing various filters convolving over the input images. Moreover, when the filters are convolved over all the data, then a feature map is formed. These feature maps are reduced in dimensions employing pooling layers minimizing the system's computational power. These feature maps can be fed again to further convolution layers by repeating the above steps.

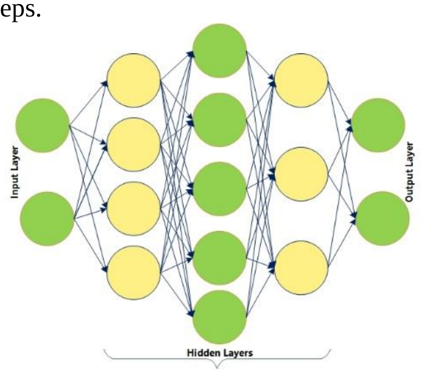

Figure 1. General Architecture of Deep Learning Model

Numerous applications[42] exist for various purposes, such as facial feature recognition [43], speech identification[35], and emotion [44]. The presented system consists of three main phases. 1) Features Extraction, 2) Training, and 3) classification. We employed features extraction utilizing a feature extraction layer through which mel-spectrograms have been generated and passed to our customized VGG19 model as the base network in YamNet [45] and two DL models i.e., ResNet50 and InceptionNetv2. Secondly, we trained an improved network over the generated mel-spectrograms belonging to two classes: Bonafide (Real) and Spoofed (Fake). The mel-spectrograms of fake audio are different from real audio. Therefore, the proposed system learns the patterns precisely for two classes. Thirdly, we classified various input audios using the trained classifier.

Then, to strengthen the process of training and classification, we introduced an ensemble learning block comprising of two DL models i.e., ResNet50, and InceptionNetV2. The detailed architecture of ensemble block is shown in Figure 3.

Undoubtedly, deep learning networks exhibit non-linear characteristics and offer valuable flexibility in situations where training datasets are limited [42]. These networks are highly sensitive to the specifics of the training data, as they are fine-tuned using random algorithms, resulting in variations in the weight sets during each training session. Consequently, neural networks can produce different predictions, leading to a high level of variance. To mitigate this variability in deep neural networks, ensemble learning has recently been employed, involving the use of multiple deep learning models instead of a single one [46]. The final prediction is then obtained by combining the predictions of these diverse models. Ensemble learning effectively merges the decisions of different models, allowing for the incorporation of more intricate and significant image features, and capturing a greater amount of useful information from various classifiers. As a result, this approach yields more reliable classification outcomes. The design of the proposed system is shown in Figure 2. In the end, we employed the voting scheme based on majority and then the model predicts the final class.

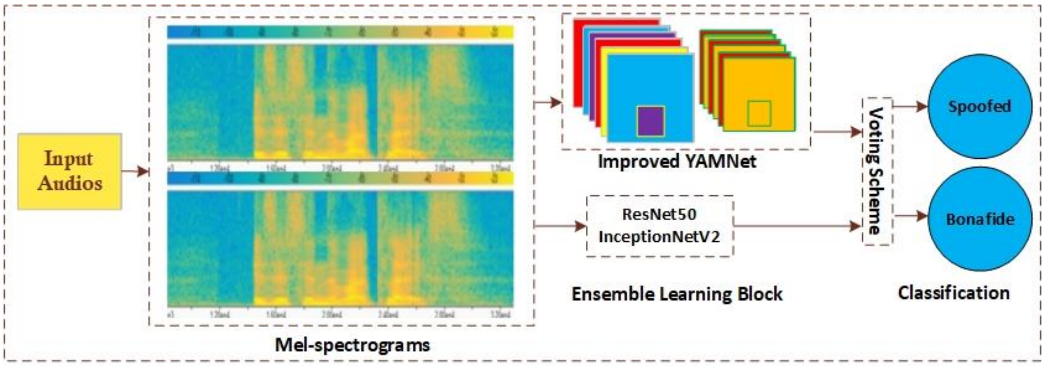

Figure 2. Flow diagram for the proposed system

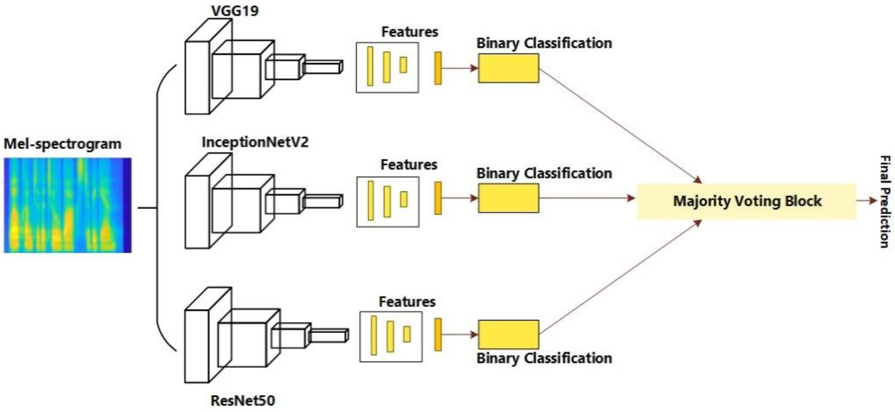

Figure 3. The process of ensemble learning and majority voting scheme

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3332561

A. YAMNET ARCHITECTURE

The key aspect of transfer learning is to minimize the computational cost by utilizing previously learned patterns. It is preferred to employ the transfer learning concept when a large size of unlabeled data is available to train a model. Therefore, the pre-trained model utilizes its previous training features to reduce the time and effort. YAMNet [45] employs the MobileNetV1 as the base network, a pre-trained model on the Google AudioSet dataset for 521 audio events. Before the features extraction phase, resampling is performed into 16000 Hz with one channel audio. Moreover, YAMNet is a DL-based model that automatically extracts audio features due to the feature extraction layer. The feature extraction layer extracts the audio features in the form of spectrograms, which are then fed to an improved MobileNet layer for classification. The layered architecture of the original YAMNet is shown in Figure 4.

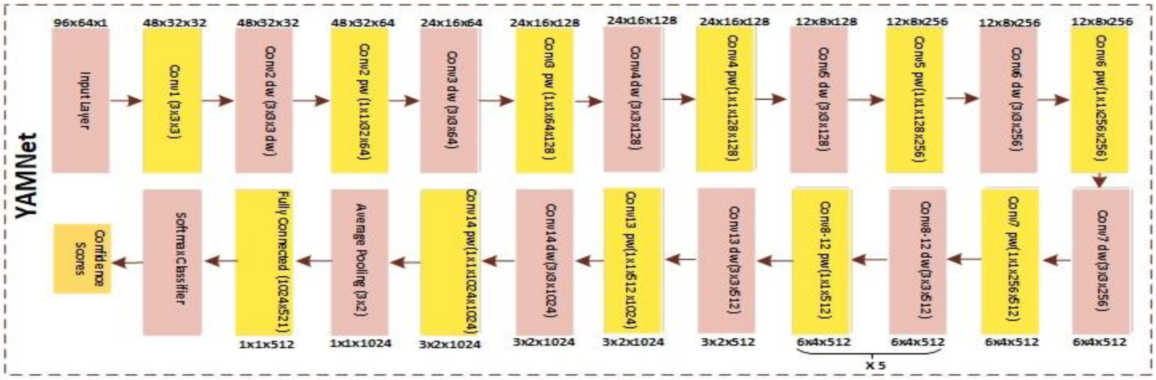

Figure 4. Architecture of original YAMNet

B. VGG19 AS BASE NETWORK

This section describes our proposed base model, i.e., visual geometry group network (VGG19) [47] architecture. The details of the layers are shown in Table I. It is a deep neural network having multi-layered functions. Initially, it was developed for the ImageNet dataset classification; then, it was found to be useful due to its structure of 3x3 convolutional operations which are stacked on the upper level and increase according to depth level. Moreover, max-pooling layers have been employed to minimize the volume size. We have replaced the convolutional layers with the grouped convolutional layers due to the size mismatch among layers input and output to utilize it for mel-spectrograms. The grouped convolutional layers are used for the features extraction and max pooling layers are utilized to decrease the dimensions of features.

More precisely, a mel-spectrogram in the form of the 2D image having dimensions of 96 x 64 x 1 is fed to the first convolutional 2d layer from an image input layer. After passing through convolution operation as 32 3x3x1 having stride 2, activation becomes 48x32x32. Then at the 3rd level, the ReLU activation function was employed. At the 4th stage, grouped convolution is applied as 32 groups of 1 3x3x1 having stride 1. The activation remains the same as 48x32x32 and passes at the 5th level through ReLU activation.

Further, at the 6th level, 5x5 max pooling is employed with stride 1, and the activation as 48x32x32 passes again from grouped convolution and ReLU activation at the 7th and 8th levels. Moreover, till the 38th level, various combinations of group convolution, max pooling, and ReLU activation have been utilized to give 48x48x32 activations. At the 39th level, a fully connected layer has been employed, giving 1x1x1024 activations with weights:1024 x 49152. At the 40th and the 41st stage, ReLU and dropout layers were used, giving 1x1x1024 activation. From 42 to 44, the same structure of fully connected, ReLU, and dropout layer is repeated. Furthermore, at the 45th stage, a dense layer is employed, giving 1x1x2 activation, which is classified through the softmax function and output is attained on the 47th layer as bonafide or spoofed. The layered architecture of proposed improved YAMNet is shown in Figure 5.

C. INCEPTIONNETV2

This network represents an enhancement of InceptionResNetv1[48] by incorporating the network structure of InceptionResNetv1 while utilizing the stem from InceptionV4. Each module within the network includes a shortcut connection on the left side. By combining the inception architecture with residual connections, the overall classification accuracy is improved. To ensure the effectiveness of the residual links, the input and output of the inception module's convolutional operation must be of the same size. Therefore, a 1x1 convolution is employed after the original convolution to match the depth dimensions. The introduction of the residual connections has led to the replacement of pooling operations. The steps for proposed system are shown below.

Algorithm 1: Steps for Proposed Spoofing Detector(EDL-Det)

```
Input: Audio samples
Output: Classified Audios as Bonafide or Spoofed
Start:
1.[$A_{train}$,$A_{test}$] ← Split Audios into train and validation sets
2.Pro_Audios ← Resampling(16000Hz, $A_{train}$ )
3.€ ← Bark-Spectrum(Pro_Audios)
4.Ms ← Mel_spec (Image_size, € // Image_size=96 x 64
5. For ∀ Ms x in → $A_{train}$            //  Start Training
    a)Image Input layer having activations of 96x64x1
    b) VGG-19, ResNet50, InceptionNetV2
    c) Majority Voting Scheme
   d) Classification
    End For
6.Ꝏ←Trained_network
7.WHILE ∀ x ∊ $A_{test}$
      a.Resampling(16000Hz)
      b.ή←Conversion into mel spectrograms (96 x 64)
      c.Features← ή
      d.Classification through trained classifier Ꝏ
   End While
8.Accuracy computation for Evaluation of Model
End
```

D. RESNET50

ResNet-50 utilizes skip connections, also called residual connections, to facilitate the learning of residual functions, enabling the network to efficiently propagate gradients during training [49]. This technique effectively addresses the degradation problem frequently encountered by deep neural networks. The structure of ResNet-50 consists of a sequence of convolutional layers, followed by multiple blocks. Each block comprises multiple convolutional layers and shortcut connections. By incorporating these shortcut connections, which bypass one or more layers, the network gains the ability to learn residual mappings. Consequently, ResNet-50 can acquire deeper representations while maintaining computational efficiency.

E. VOTING SCHEME

After attaining the predictions from three DL models, we utilized majority voting scheme. The majority voting scheme is a straightforward and popular method for classification problems, particularly when used in combination with ensemble learning and vote-based algorithms. The three classifiers such as VGG-19, ResNet50, and InceptionNetV2 provided individual classification results either as spoofed or bonafide. Then, in voting scheme block, the class with higher votes is selected as output class.

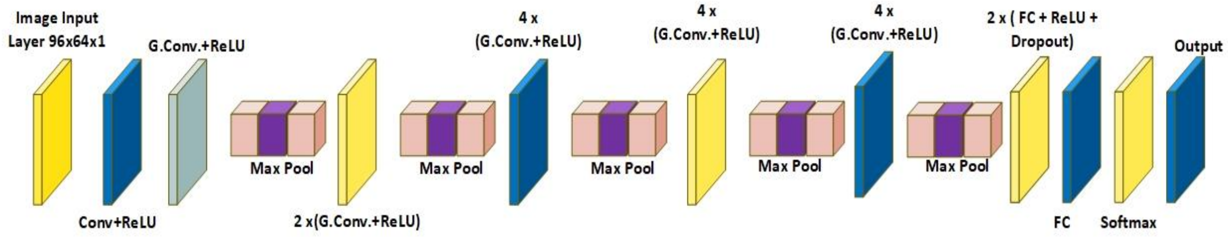

Figure 5. Layered architecture of the proposed modified YAMNet (FC: Fully Connected Layer, G.Conv.: Grouped Convolution)

Table I: LAYER-WISE DETAILS OF VGG19 BASED YAMNET

Type	Activations	Learnable	Stride/Channel	Total Learnable
Image Input	96 x 64 x 1	-	-	0
Convolution 2D (Conv)	48 x 32 x 32	Weights: 3 x 3 x 1 x 32 Bias: 1 x 1 x 32	32 3 x 3 x 1 convolutions Stride: [2 2] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0

MaxPool	48 x 32 x 32	-	5x5 max pooling Stride: [1 1] Padding: Same	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
MaxPool	48 x 32 x 32	-	5x5 max pooling Stride: [1 1] Padding: Same	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
MaxPool	48 x 32 x 32	-	5x5 max pooling Stride: [1 1] Padding: Same	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions	320

		Bias: 1 x 1 x 1 x 32	Stride: [1 1] Padding: Same	
ReLU	48 x 32 x 32	-	-	0
Grouped Convolution	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
ReLU	48 x 32 x 32	-	-	0
MaxPool	48 x 32 x 32	-	5x5 max pooling Stride: [1 1] Padding: Same	0
4 x (Grouped Convolution+ ReLU)	48 x 32 x 32	Weights: 3 x 3 x 1 x 1 x 32 Bias: 1 x 1 x 1 x 32	32 groups of 1 3x3x1 convolutions Stride: [1 1] Padding: Same	320
	48 x 32 x 32	-	-	0
MaxPool	48 x 32 x 32	-	5x5 max pooling Stride: [1 1] Padding: Same	0
2x[Fully Connected Layer + ReLU + Dropout]	1 x 1 x 1024 1 x 1 x 1024 1 x 1 x 1024	Weights:1024x1024 Bias: 1024x1	1024 fully connected layer	50332672, 1049600
Fully Connected Layer	1x1x2	Weights:2x1024 Bias: 2x1	2 fully connected layer	2050
Softmax	1x1x2	-	Softmax	0
Classification	1x1x2	-	Output	0

IV. EXPERIMENTAL EVALUATION

A. ENVIRONMENTAL SETUP

For the experiments, we used s system having a GPU NVIDIA card, GEFORCE GTX(4GB). The details of the utilized system are reported in Table II. The experiments were accomplished using Matlab 2021a.

TABLE II: SYSTEM SPECIFICATIONS FOR THE EMPLOYED MODEL

Hardware	Specifications
Graphical Processing Unit	NVIDIA GEFORCE GTX x 4
Computer	GPU Server
Central Processing Unit	Intel Core i5

B. METRICS

To evaluate the working of EDL-Det, we employed several metrics, including Precision, Recall, Accuracy, Equal Error Rate, and the Tandem-Detection Cost Function (t-DCF). The precision is the ratio of a number of positives to the total number of audio samples (mel-spectrograms) categorized as spoofed. The equation for precision is provided below.

$$Precision = \frac{TP}{TP+FP}, \tag{1}$$

The accuracy of the detector refers to the proportion of correctly categorized audio by EDL-Det. The corresponding equation is provided below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

The recall is the proportion of audios from the positive class that was correctly categorized by the proposed detector among all the spoofed audios, including those classified as genuine by the model. A higher recall number, closer to 1, indicates a better-performing model. The mathematical equation for the recall is presented below.

$$Recall = \frac{TP}{TP+FN}, \tag{3}$$

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3332561

Moreover, we utilized Equal Error Rate (EER) and t-DCF to analyze the working of the proposed TTS synthesis detector.

C. EXPERIMENTAL PROTOCOLS

In this segment, we define the protocols of executed experiments to assess the performance of the suggested model. To evaluate the model on the LA set, we utilized the training samples as 25,380, including 2580 bonafide samples and 22800 spoofed samples, to train our spoofing detector. We performed testing of EDL-Det using both sets, such as eval and dev sets. The eval set consists of 71,237 samples, including 7355 spoofed and 63882 bonafide samples, while the dev set consists of 24,844 samples, including 22,296 spoofed and 2548 bonafide samples. Furthermore, we have evaluated our proposed model using the PA dataset. We employed 54,000 samples, including 48,600 spoofed and 5400 bonafide samples, for the model's training. We also evaluated EDL-Det using both the remaining sets, that is, eval and dev sets. The eval set comprises 1,34,730 samples, including 116,640 spoofed and 18,090 bonafide samples and the dev set comprises 29,700 audios, including 24,300 spoofed and 5400 bonafide samples.

D. DATASET

The contest of spoofed voice detection came in 2015, known as ASVSpoof 2015 corpus [50]. The aim was to develop a system to detect the synthesized or cloned speech and analyze the performance using the dataset samples. After two years, ASVSpoof 2017 corpus [28] came into existence to evaluate the replay detection systems. A large and assorted dataset was introduced in 2019, known as ASVSpoof 2019 [51], comprising both logical and physical access attacks. It was split into two parts such as LA and PA. The first contained the voice conversion and synthesized speech samples, including bonafide audio. The later part consists of replay and bonafide audio samples. Furthermore, both parts have been split further into 3 sub-parts: development, training, and evaluation sets. The LA dataset consists of 17 TTS and voice cloning systems. Moreover, these systems are trained using the voice cloning toolkit VCTK [52].Among these systems, 6 have been labeled as known attacks, whereas the other 11 systems are known as anonymous attacks. The training and dev. audio samples are taken from known attacking systems, and evaluation samples are collected from 11 unknown and 2 known attacks. The Logical Access set consists of 2 VC systems that utilize spectral filter and artificial neural networks-based approaches. Furthermore, the LA set consists of 4 TTS systems that utilize artificial neural networks or waveform concatenation employing vocoders based on source-based filter Vocoder [53], or WaveNet Vocoder [54]. The 11 unknown spoofing methods consist of 2 VC, 6 TTS, and 3 Hybrid forms of VC and TTS systems utilizing various waveform-based methods such as GriffinLim [55], Neural waveform techniques [56], Generative adversarial networks (GAN) [57], and combinations of waveform and spectral filtering. The indicators of the ASVSpoof 2019 dataset are reported in Table III, whereas the in-depth summary of the LA set is shown in Table IV. Moreover, ASVspoof 2017 [28] comprises real replay speeches, while ASVspoof 2019 comprises synthesized replay recordings recorded under an acoustic environment to enrich the ASV system's reliability. Training and development (dev.) recordings are produced, conferring to 9 replay and 27 acoustic configurations. The sizes of rooms are categorized as large, medium, and small rooms. All speeches are generated in various zones, such as A, B, and C, exhibiting varying distances (Da) between the talker and zone. The zone A voice quality is better than B and C zone. Moreover, the eval recordings have been gathered in the same way as train and dev sets.

TABLE III:STATISTICS OF ASVSPOOF 2019 LA AND PA SETS

Set	LA		PA	
	Spoofed	Bonafide	Spoofed	Bonafide
Train	22800	2580	48600	5400
Evaluate	63882	7355	116640	18090
Dev.	22296	2548	24300	5400
Total	36326	12483	189540	28890

TABLE IV: SUMMARY OF ASVSPOOF 2019 LA SPOOFING SYSTEMS

Label	Input	Conversion	Outputs	Processor	Post Process	Speaker Representation	Waveform Generator
A01	Text	AR Recurrent Neural Network	F0 MCC	Natural Language Processing	NA	Variational Auto-Encoder(VAE)	WaveNet

A02	Text	AR Recurrent Neural Network	F0 MCC BAP	Natural Language Processing	NA	VAE	WORLD
A03	Text	Feed Forward Neural Network	F0 MCC BAP	Natural Language Processing	NA	Single hot embedding	WORLD
A04	Text	CART	F0 MCC	Natural Language Processor	NA	NA	Waveform Concatenation
A05	Human Speech	VAE Neural Network	F0 MCC AP	WORLD	NA	Single hot embedding	WORLD
A06	Human Speech	GMM-UBM	LPC	MFCC/LPCC	NA	NA	OLA and Special Filters
A07	Human Speech	Recurrent Neural Network	F0 MCC BA	Natural Language Processor	Generative Adversarial Network	Single hot embedding	WORLD
A08	Human Speech	AR Recurrent Neural Network	F0 MCC	Natural Language Processor	NA	Single hot embedding	Neural Source Filter Neural Network
A09	Human Speech	Recurrent Neural Network	F0 MCC	Natural Language Processor	NA	Single hot embedding	Vocaine
A10	Human Speech	AR Recurrent and CNN	Spectrograms	CNN and Bi RNN	NA	Recurrent Neural Network(d vector)	Wave Recurrent Neural Network
A11	Human Speech	AR Recurrent and CNN	Spectrograms	CNN and Bi RNN	NA	Recurrent Neural Network(d vector)	Griffin-Lim
A12	Human Speech	Recurrent Neural Network	Linguistic-based features and F0	Natural Language Processor	NA	Single hot embedding	WaveNet Neural Network
A13	TTS	Moment Match	MCC	WORLD	NA	NA	Waveform based filtering
A14	TTS	Recurrent Neural Network	F0 MCC BAP	ASR Neural Network	NA	NA	STRAIGHT
A15	TTS	Recurrent Neural Network	MCC F0	ASR Neural Network	NA	NA	WaveNet Network
A16	Text	CART Neural Network	F0 MFCC	Natural Language Processor	LA-A04	NA	Waveform Concatenation
A17	Human Speech	VAE Neural Network	F0 MCC	WORLD	NA	Single hot embedding	Waveform based filters
A18	Human Speech	Linear	MFCC	i-Vector/MFCC	NA	PLDA	MFCC Vocoder
A19	Human Speech	GMM-UBM	LPC	MFCC/LPCC	LA-A06	NA	OLA and Special filters

E. SYNTHESIZED SPEECH AND VOICE CONVERSION DETECTION

This section will evaluate EDL-Det over text-to-speech synthesis (TTS) and voice conversions (VC) samples. Therefore, we employed three DL methods to classify the speeches into bonafide and spoofed speeches of TTS and voice conversion samples. In the datasets, 4 TTS spoofed systems exist, including A01, A02, A03, and A04, whereas 2 VC spoofed methods, including A05 and A06, are utilized to generate spoofed samples for the LA dataset for training. In the eval set of LA, 13 spoofed systems are included comprising 7 text-to-speech syntheses: A07-A12, A16, 3 TTS-VC systems; A13, A14, A15, and 3 VC spoofed systems; A17-A19 that are used to generate the spoof speeches. We employed an experiment based on three phases to assess the efficacy of EDL-Det for VC and TTS systems. The mel-spectrograms for real and fake audios are shown in Figure 6. From the figure, it is clearly visible that when there is pause in audio, the corresponding region becomes entirely blank in fake audio, however it contains some patterns due to the presence of noise. The vertical yellow patterns in real audio represent the pitch and emphasize on the words. Moreover, the reason of non-linear patterns in mel-spectrogram of real audio is the high background noise. Whereas, the patterns are less varying for fake audio due to the same pitch throughout the full audio.

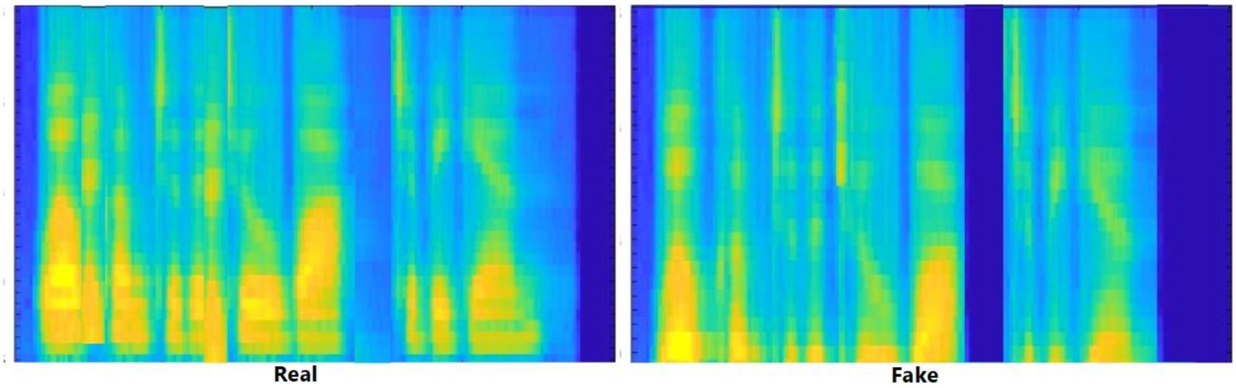

Figure 6. The visual appearance of audios as mel-spectrograms

The statistics are shown in Table V. We attained an EER of 0.51% and t-DCF of 0.005 over an eval set of LA dataset. Second, we utilized samples from the VC system of the train set from the LA dataset for training. We attained an EER of 11.01% and t-DCF of 0.07 on the eval set of VC from LA samples. The statistics are presented in Table V. It is concluded from the outcomes that the suggested model achieves better the detection of TTS spoofed speeches than VC spoofed detection. The reason behind the better performance of the proposed model for TTS spoof detection is that the voice generated from the VC systems is based on the original audio samples' periodic characteristics. However, TTS systems lack this property. Third, we performed an experiment using the general LA dataset to analyze the working of the proposed technique and achieved 0.045 EER. The overall working of the suggested system is significant on the LA set; therefore, we can say that our model effectively detects the spoofed audios.

TABLE V: RESULTS FOR SYNTHESIZED SPEECH AND VOICE CONVERSION

Spoofing System	Accuracy (%)	Precision (%)	Recall (%)	EER (%)	Min-tDCF
VC	84.2%	96.7%	84%	1.01	0.07
TTS	99.7%	99.2%	99.5%	0.51	0.005
LA(overall)	99.7%	99.1%	99.5%	0.045	0.002

F.EVALUATION OF VOICE CLONING ALGORITHMS

In this experimentation, we aim to describe the system type employed for synthesizing spoofed audios for the ASVspoof 2019 LA dataset. LA set consists of samples produced employing TTS and voice conversion systems. More precisely, 6 types of cloning algorithms have been utilized for speech synthesis in the LA set of ASVspoof 2019. These algorithms include the TTS vocoder, TTS waveform model, TTS waveform concatenation, and spectral filtering. We utilized 22,800 samples from the LA collection for EDL-Det's training and the dev samples (22,296 samples) for testing. We achieved an EER of 0.8%, 1.2%, 0.5%, 1.9%, 1.4%, and 2.5% for A01-A06 respectively. The statistics are reported in Table VI, and it can be analyzed that the proposed technique provides the best performance over the A03 algorithm, which is a vocoder and utilizes the WORLD mechanism for waveform generation. Moreover, we achieved the lowest accuracies for A04 and A06. A04 is based on waveform concatenation for speech generation, and A06 is based on spectral filtering and employs OLA along with spectral filtering for speech generation. Therefore, it is observable from the experiment that our proposed model is slightly less operative in analyzing the A04 and A06 algorithms than other algorithms. However, on average, we achieved excellent results for cloning-based speech detection. Therefore, we can conclude that the cloning-based speeches with their artifacts can be distinguished easily through our proposed system, for the feature extraction and classification of mel spectrograms. The capability of detecting the cloning-based speeches of our proposed model makes it more effective and significant for applications in audio forensics.

TABLE VI: PERFORMANCE OF EDL-DET OVER CLONING ALGORITHMS

Algorithm	Accuracy(%)	Precision(%)	Recall(%)	EER(%)
A01(TTS neural waveform)	97.7	97.9	97.9	0.8
A02(TTS vocoder)	96.9	93.7	97.2	1.2
A03(TTS vocoder)	99.1	98.6	98.3	0.5
A04(TTS waveform concatenation)	94.8	94.2	95.6	1.9
A05(VC vocoder)	97.2	97.8	98.2	1.4
A06(VC spectral)	94.7	95.1	96.3	2.5

G.EVALUATION OF LA AND PA ATTACKS

In this unit, we target to examine the working of our spoofing audio detector using LA and PA attacks. Therefore, we transformed the auditory samples of LA and PA set into mel-spectrograms and passed them to an improved YAMNet's base network, VGG-19, ResNet50, and InceptionNetv2 for the classification into bonafide and spoofed. For LA attacks, we achieved an EER of 0.80%, and 0.045% for Dev and Eval sets, respectively. Moreover, we achieved tDCF of 0.04 and 0.002 for Dev and Eval, respectively. From the results, we believe that the working of the suggested model is enhanced than the existing spoofing detectors for LA attacks [58]. Moreover, for PA attacks, we achieved an EER of 0.48% and 3.2% for eval and dev sets. The min-tDCF of 0.003 and 0.06 is attained for eval and dev sets, as reported in Table VII.

It can be examined from the results that our suggested spoofing detector attained a significant performance than the existing models. Remarkably, our proposed model is based on three DL methods which utilizes the combined effect of convolutional layers to extract the most representative features from the mel-spectrograms generated from audios. Therefore, we attained 0.045% and 0.48% EER, which is less than the EER achieved for the existing system on the eval set, such as in [58]. We believe, after the experiment, that EDL-Det can effectively extract features from replay samples to identify physical access attacks.

TABLE VII: RESULTS ON LA AND PA SET OF ASVSPOOF 2019

Set	Set Category	Accuracy (%)	Precision (%)	Recall (%)	EER (%)	Min-tDCF
LA	Dev	99.4	99.9	99.4	0.80	0.04
	Eval	99.7	99.1	99.5	0.045	0.002
PA	Dev	99.2	98.1	98.3	3.20	0.06
	Eval	99.8	99.2	98.4	0.48	0.003

H.COMPARATIVE ANALYSIS WITH EXISTING FEATURES EXTRACTION-BASED TECHNIQUES

In this section, we experimented to compare our spoofing detector with the present models for spoofing voice detection based on hand-crafted feature extraction. To validate the proposed model's efficacy for detecting artifacts of cloned voices, detection of replay distortions, and prosodic features of the speech, we employ a comparative analysis with the state-of-the-art methods, as presented in Table VIII. We analyzed the working of the proposed technique using LA and PA Eval sets of the ASVspoof 2019 dataset employing t-DCF and EER metrics. For the LA set, the best EER is 0.045%, and our proposed spoofing detector attains a t-DCF of 0.002. The second-best EER is 0.06, and t-DCF is 0.0017 attained by [59]. Furthermore, 9.33%, 7.69%, 8.09%, 9.57%, and 2.502% EER are achieved by MFCC-ResNet[13], CQCC-ResNet [13], LFCC-GMM [58], and CQCC-GMM [58] respectively. Similarly, another model [59] existing techniques in terms of accuracy, EER, and t - DCF.

attained 0.58% EER and 0.0160 t-DCF for the PA eval set. Other algorithms, such as [13] and [58], performed various experiments and attained 4.43%, 13.54%, 1.04%, and 0.459% EER. Furthermore, for the PA eval set, the best EER is 0.48, and t-DCF is 0.003 attained by our proposed model. Moreover, the second-best EER is 0.459%, and t-DCF of 0.0116 was achieved by [60]. From this analysis, it is concluded that our spoofing detector can detect various spoofed attacks and voices based on cloning algorithms effectively. More precisely, our proposed algorithm outperforms the

TABLE VIII: COMPARISON WITH EXISTING SPOOFING DETECTION SYSTEMS

Model	LA (Eval Set)		PA (Eval Set)	
	t-DCF	EER(%)	Min-tDCF	EER(%)
MFCC-ResNet [13]	0.2042	9.33	-	-
CQCC-ResNet [13]	0.2166	7.69	0.1070	4.43
Baseline LFCC+GMM [58]	0.2116	8.09	0.3017	13.54
Baseline CQCC+GMM [58]	0.2366	9.57	0.2454	11.04
CQT+CE+ SE_ResNet50 [60]	0.0743	2.502	0.0116	0.459
CLS-LBP+LSTM [59]	0.0017	0.06	0.0160	0.58
Proposed Model	**0.002**	**0.045**	**0.003**	**0.48**

I. COMPARISON WITH EXISTING DL MODELS BASED ON ACCURACY

In this section, we evaluate our proposed system with existing DL-based methods for fake audio detection. In [61], a CNN has been proposed using data augmentation and dropout to avoid overfitting. The method attained 99.7% recall, 99.7% recall, and 98.5% accuracy overall. Wijethunga et al. [62] developed a technique for fake audio detection employing four main steps. The four steps included audio de-noising for the pre-processing of speeches, speaker diarization for the conversion of text, and RNN for the labeling of the speaker. Steven et al. [63] developed a system based on customized CNN and evaluated the model using FoR dataset consisting of various synthesized audios employing deep fake generation techniques. The model attained 88.9% classification accuracy. Janavi et al. [64] proposed various algorithms for classifying fake audios, such as SVM, KNN, RF, etc. and a temporal convolutional network (TCN). The test data results showed that the TCN model achieved 92% accuracy, comparatively higher than the machine learning-based techniques. In the end, classification was performed, attaining an accuracy of 94%. However, our proposed TTS detector is also robust and attains 99.7% accuracy over the LA set. Therefore, the comparison ensures that our proposed TTS detector is better than existing deep learning-based models.

J. ROBUSTNESS

To evaluate the robustness of our model over unseen attacks, we experimented on diverse and large-scale ASVspoof 2019 samples. It is worth mentioning that the dataset comprises 87 unseen speakers samples that were utilized for the evaluation purpose, whereas for training, samples from 20 speakers were used. Moreover, spoofed voices utilized in the training set have been cloned employing 6 algorithms, whereas spoofed voices utilized for the evaluation set were cloned using 19 cloning techniques employing 13 advanced algorithms.

Therefore, we assessed the performance to recognize the unseen spoofing attacks that are synthesized using complex spoofing techniques such as A07-A19. It is also considered that the ASVspoof's eval and train sets consist of audio samples from various speakers, multiple algorithms for LA attacks such as voice conversion and TTS, and variations in background environment and microphones for PA attacks. Thus, it is concluded that the PA and LA evaluation sets have more diverse conditions than the training sets. It can be assessed that our proposed spoofing detector achieves excellent results over the evaluation set that is more diverse comprising unseen attacks and speakers, varying environments, and microphones than training sets. The variations of the evaluation set did not affect the performance, which is evidence of the robustness of our proposed technique as exhibited in Tables V, VI, and VII Furthermore, our proposed model can effectively detect all types of LA and PA attacks, such as voice conversion, replay attacks, and text-to-speech.

K. ABLATION STUDY

In this section, we are performing two experiments to analyze the efficacy of EDL-Det. In the first experiment, we analyze the results of our proposed detector, VGG-19, ResNet50, InceptionNetV2, and original YAMNet with MobileNet as the base network. For this purpose, we used TTS samples of training and eval sets from the LA subset to train and test respectively. We used hyperparameters as: learning rate:0.001, batch size:64, epochs:150, and a number of iterations:1000. The results are reported in Table IX. The base network achieves 2.3% EER, 92.1% accuracy, 93.4% precision, and 94.1% recall. Whereas, our proposed detector EDL-Det achieves 0.045% EER, 99.7% accuracy, 99.1% precision, and 99.5% recall. Moreover, the individual models i.e., VGG19, ResNet50, and InceptionNetv2 attains considerable results, however, when they are combined, they provide the promising results. Therefore, It is clearly visible that our proposed detector attains significant results than the base network under the same environment.

TABLE IX: PERFORMANCE OF THE PROPOSED DETECTOR VS BASE MODEL

Model	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**EER(%)**
Base Network	92.1	93.4	94.1	2.3
VGG19	96.4	95.3	94.4	1.10
ResNet50	97.2	96.2	95.3	1.01
InceptionNetv2	98.5	97.4	96.3	0.92
EDL-Det	**99.7**	**99.1**	**99.5**	**0.045**

In the second experiment, we used ASVspoof 2021 dataset for the analysis of robustness by our proposed model than the original YAMNet. The hyperparameters are similar to the first experiment. The ASVspoof 2021 comprises three subsets: LA, PA, and speech deepfake. We used the samples from a speech deepfake set that involves audio through a combination of genuine and manipulated speeches generated using TTS and VC algorithms. It is similar to the LA task, which involves compressed data, but does not require speaker verification. The results are reported in Table X. It is visible from the results that our improved model attains better results than the original YAMNet. The results also ensure the robustness of our proposed detector. The reason behind the better results is due to a fruitful effect of three DL methods.

L.COMPARISON WITH EXISTING DETECTORS

In this experiment, we compared EDL-Det with the existing models for voice spoofing detection. The comparative results are reported in Table XI, considering the evaluation set of the ASVspoof 2019 LA corpus. Our proposed spoofing detector attains the lowest EER as 0.045 on eval sets and outperforms the existing systems. On the other hand, W2V2-light-DARTS achieved the second lowest EER for the eval set as 1.08. In [65], silence and dual-band fusion on neural network has been employed for detection attaining 1.14 EER. Similarly, [66], [67], [68], and [69] achieved 1.87, 4.87, 1.12, and 1.15 EER respectively. From this analysis, it is concluded that our proposed TTS detector identifies spoofed speeches effectively. Additionally, our proposed TTS detector is more robust than these spoofed audios detector. The reason behind the exceptional performance is the implication of ensemble learning. It is known that ensemble models provide better generalization which is an existing issue in state-of-the-art methods. Further, the proposed EDL-Det also reduces overfitting due to combination of three DL models as individual models may overfit to different parts of data. Hence, the proposed model outperforms the existing techniques of spoofing detection specifically TTS synthesis.

TABLE XI: PERFORMANCE COMPARISON WITH EXISTING SPOOFING DETECTORS BASED ON EER

Model	EER on LA(Eval Set)
FFT-L-SENet[65]	1.14
Attention-based CNN[66]	1.87
LFCC-PC-DARTS[67]	4.87
RAWNet2[68]	1.12
W2V-Siamese[69]	1.15
W2V2-light-DARTS[70]	1.08
EDL-Det	**0.045**

TABLE X: PERFORMANCE OF THE PROPOSED DETECTOR VS BASE MODEL USING THE ASVSPOOF 2021 DATASET

Model	Accuracy(%)	Precision(%)	Recall(%)	EER(%)
Base Network	88.3	90.4	89.4	2.4
EDL-Det	**98.7**	**98.8**	**97.3**	**0.056**

V. DISCUSSION

In this study, we propose a robust TTS detector using three deep learning models in ensemble manner. The proposed EDL-Det transformed the audios into the Mel-spectrograms and then passed them for features extraction in three separate DL models. Each model might be overfit on various parts of data, however due to ensemble technique the proposed model overcame the issue of overfitting. For ensemble learning, we selected VGG19, ResNet50, and InceptionNetv2 due to their simple architecture and contribution towards better classification results. Further, the EDL-Det employs voting scheme for the classification into spoofed and bonafide category.

The issues in existing systems include lack of generalization, explainability, challenge of data imbalance, and dependence of performance evaluation on only accuracy. We overcame the issues of generalization, explanability, and utilized several standard metrics for the performance evaluation of the proposed EDL-Det. However, there exist some limitations in the proposed system such as the time required for training, and high computational cost. The proposed system could be trained fast and deployed in real-world devices with expensive computational resources.

VI. CONCLUSION

This paper presents a voice spoofing detector employing an improved deep learning model, YAMNet alongwith an ensemble learning block to detect PA and LA attacks. We employed a customized network, VGG-19, as a base network in YAMNet, and two DL models i.e., ResNet50, and InceptionNetV2 for feature extraction and classification of mel-spectrograms from bonafide and spoofed samples. A customized Vgg-19 effectively captures the sample dynamics, artifacts of cloning algorithms and environment, and microphones variations of the replay attacks. Moreover, ensemble learning block makes our proposed model more reliable and effective for classification. We assessed the performance of the proposed model using a diverse and large-scale dataset, ASVspoof 2019 corpus, and it is concluded that our proposed EDL-Det is applicable for detecting several types of spoofing attacks. More precisely, our model correspondingly attained an EER of 0.48% and 0.045% for PA and LA attacks. Our system effectively distinguishes the various cloning algorithms employed for speech generation. Additionally, our comparative assessment with existing models unveils that our proposed spoofing detector outperforms them for various forms of speech spoofing detection, such as cloning-based, text-to-speech, and replay attacks. Furthermore, it is worth mentioning that our proposed detector attained significant outcomes on ASVspoof 2021 dataset. Therefore, our proposed model is a robust spoofing detector due to its effectiveness in cross-validation over the evaluation set of ASVspoof 2019.

The one challenge that we want to overcome is to reduce the training time of our proposed detector for achieving significant performance. Moreover, we aim to cross-validate our model on other voice spoofing datasets and further improve the performance.

ACNKNOWLEDGEMENT: This research work was supported by University of Engineering and Technology Taxila, Pakistan.

REFERENCES

[1] Mittal, A. and M. Dua, Automatic speaker verification systems and spoof detection techniques: review and analysis. International Journal of Speech Technology, 2022: p. 1-30.

[2] Cho, Y.P., et al. Mandarin Singing Voice Synthesis with Denoising Diffusion Probabilistic Wasserstein GAN. in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2022. IEEE.

[3] García, V., I. Hernáez, and E. Navas, Evaluation of Tacotron Based Synthesizers for Spanish and Basque. Applied Sciences, 2022. 12(3): p. 1686.

[4] Săracu, G. and A. Stan. An analysis of the data efficiency in Tacotron2 speech synthesis system. in 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). 2021. IEEE.

[5] Oord, A., et al. Parallel wavenet: Fast high-fidelity speech synthesis. in International conference on machine learning. 2018. PMLR.

[6] Mirsky, Y. and W. Lee, The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 2021. 54(1): p. 1-41.

[7] Ahmed, S.R., et al. Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks. in 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). 2022. IEEE.

[8] Almutairi, Z. and H. Elgibreen, A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms, 2022. 15(5): p. 155.

[9] Todisco, M., et al., ASVspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441, 2019.

[10] Dinkel, H., 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, 2018. 26(11): p. 2002-2014.

[11] Chintha, A., et al., Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE Journal of Selected Topics in Signal Processing, 2020. 14(5): p. 1024-1037.

[12] Ma, X., et al. Improved lightcnn with attention modules for asv spoofing detection. in 2021 IEEE International Conference on Multimedia and Expo (ICME). 2021. IEEE.

[13] Alzantot, M., Z. Wang, and M.B. Srivastava, Deep residual neural networks for audio spoofing detection. arXiv preprint arXiv:1907.00501, 2019.

[14] Hamza, A., et al., Deepfake Audio Detection via MFCC Features Using Machine Learning. IEEE Access, 2022. 10: p. 134018-134028.

[15] Salman, S., J.A. Shamsi, and R. Qureshi, Deep Fake Generation and Detection: Issues, Challenges, and Solutions. IT Professional, 2023. 25(1): p. 52-59.

[16] Stroebel, L., et al., A systematic literature review on the effectiveness of deepfake detection techniques. Journal of Cyber Security Technology, 2023: p. 1-31.

[17] Chao, Y.-H., et al., Using kernel discriminant analysis to improve the characterization of the alternative hypothesis for speaker verification. IEEE transactions on audio, speech, and language processing, 2008. 16(8): p. 1675-1684.

[18] Ze, H., A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. in 2013 ieee international conference on acoustics, speech and signal processing. 2013. IEEE.

[19] Dörfler, M., R. Bammer, and T. Grill. Inside the spectrogram: Convolutional Neural Networks in audio processing. in 2017 international conference on sampling theory and applications (SampTA). 2017. IEEE.

[20] Balamurali, B., et al., Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. IEEE Access, 2019. 7: p. 84229-84241.

[21] Sharma, H.K., et al. CNN-Based Model for Deepfake Video and Image Identification Using GAN. in Proceedings of Fourth International Conference on Computer and Communication Technologies. 2023. Springer.

[22] Wu, Z., 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, 2016. 24(4): p. 768-783.

[23] Yaman, S. and J. Pelecanos, Using polynomial kernel support vector machines for speaker verification. IEEE Signal Processing Letters, 2013. 20(9): p. 901-904.

[24] Loughran, R., et al., Feature selection for speaker verification using genetic programming. Evolutionary Intelligence, 2017. 10(1): p. 1-21.

[25] Zhao, H. and H. Malik, Audio recording location identification using acoustic environment signature. IEEE Transactions on Information Forensics and Security, 2013. 8(11): p. 1746-1759.

[26] AlBadawy, E.A., S. Lyu, and H. Farid. Detecting AI-Synthesized Speech Using Bispectral Analysis. in CVPR Workshops. 2019.

[27] Paul, D., M. Pal, and G. Saha, Spectral features for synthetic speech detection. IEEE journal of selected topics in signal processing, 2017. 11(4): p. 605-617.

[28] Kinnunen, T., et al., The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.

[29] Yu, H., et al., Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features. IEEE transactions on neural networks and learning systems, 2017. 29(10): p. 4633-4644.

[30] Maccagno, A., et al., A CNN approach for audio classification in construction sites, in Progresses in Artificial Intelligence and Neural Systems. 2021, Springer. p. 371-381.

[31] Bai, S., J.Z. Kolter, and V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.

[32] Zhang, C., C. Yu, and J.H. Hansen, An investigation of deep-learning frameworks for speaker verification antispoofing. IEEE Journal of Selected Topics in Signal Processing, 2017. 11(4): p. 684-694.

[33] Luo, A., et al. A capsule network based approach for detection of audio spoofing attacks. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.

[34] Wang, R., et al. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. in Proceedings of the 28th ACM International Conference on Multimedia. 2020.

[35] Dhamyal, H., et al., Fake Audio Detection in Resource-constrained Settings using Microfeatures. Proc. Interspeech 2021, 2021: p. 4149-4153.

[36] Lea, C., et al. Temporal convolutional networks: A unified approach to action segmentation. in European Conference on Computer Vision. 2016. Springer.

[37] Arık, S.Ö., et al. Deep voice: Real-time neural text-to-speech. in International Conference on Machine Learning. 2017. PMLR.

[38] Ping, W., et al., Deep Voice 3: 2000-Speaker Neural Text-to-Speech. 2017.

[39] Ballesteros L, D.M. and J.M. Moreno A, Highly transparent steganography model of speech signals using Efficient Wavelet Masking. Expert Systems with Applications, 2012. 39(10): p. 9141-9149.

[40] Ballesteros L, D.M. and J.M. Moreno A, On the ability of adaptation of speech signals and data hiding. Expert Systems with Applications, 2012. 39(16): p. 12574-12579.

[41] Liu, T., et al., Identification of Fake Stereo Audio Using SVM and CNN. Information, 2021. 12(7): p. 263.

[42] Mahum, R. and A. AlSalman, Lung-RetinaNet: Lung Cancer Detection using a RetinaNet with Multi-Scale Feature Fusion and Context Module. IEEE Access, 2023.

[43] Yang, S., et al. From facial parts responses to face detection: A deep learning approach. in Proceedings of the IEEE international conference on computer vision. 2015.

[44] Ng, H.-W., et al. Deep learning for emotion recognition on small datasets using transfer learning. in Proceedings of the 2015 ACM on international conference on multimodal interaction. 2015.

[45] M. Plakal and D. Ellis, Y., " . [Online]. Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, [Online]. Available:. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, Jan 2020.

[46] Yazdinejad, A., et al., An ensemble deep learning model for cyber threat hunting in industrial internet of things. Digital Communications and Networks, 2023. 9(1): p. 101-110.

[47] Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[48] Szegedy, C., et al. Inception-v4, inception-resnet and the impact of residual connections on learning. in Proceedings of the AAAI conference on artificial intelligence. 2017.

[49] He, K., et al. Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[50] Wu, Z., et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. in Sixteenth annual conference of the international speech communication association. 2015.

[51] Wang, X., et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language, 2020. 64: p. 101114.

[52] Veaux, C., J. Yamagishi, and K. MacDonald, Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.

[53] Morise, M., F. Yokomori, and K. Ozawa, World: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE TRANSACTIONS on Information and Systems, 2016. 99(7): p. 1877-1884.

[54] Oord, A.v.d., et al., Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[55] Griffin, D. and J. Lim, Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 1984. 32(2): p. 236-243.

[56] Wang, X., S. Takaki, and J. Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. IEEE.

[57] Tanaka, K., et al. Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks. in 2018 IEEE Spoken Language Technology Workshop (SLT). 2018. IEEE.

[58] Todisco, M., H. Delgado, and N. Evans, Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language, 2017. 45: p. 516-535.

[59] Dawood, H., et al., A robust voice spoofing detection system using novel CLS-LBP features and LSTM. Journal of King Saud University-Computer and Information Sciences, 2022. 34(9): p. 7300-7312.

[60] Li, X., et al. Replay and synthetic speech detection with res2net architecture. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.

[61] Ballesteros, D.M., et al., Deep4SNet: deep learning for fake speech classification. Expert Systems with Applications, 2021. 184: p. 115465.

[62] Wijethunga, R., et al. Deepfake audio detection: a deep learning based solution for group conversations. in 2020 2nd International Conference on Advancements in Computing (ICAC). 2020. IEEE.

[63] Camacho, S., D.M. Ballesteros, and D. Renza. Fake speech recognition using deep learning. in Workshop on Engineering Applications. 2021. Springer.

[64] Khochare, J., et al., A deep learning framework for audio deepfake detection. Arabian Journal for Science and Engineering, 2022. 47(3): p. 3447-3458

[65] Zhang12, Y., W. Wang12, and P. Zhang12, The effect of silence and dual-band fusion in anti-spoofing system. 2021.

[66] Ling, H., et al. *Attention-Based Convolutional Neural Network for ASV Spoofing Detection*. in *Interspeech*. 2021.

[67] Liu, S., et al., *Recent progress in the CUHK dysarthric speech recognition system.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. **29**: p. 2267-2281.

[68] Tak, H., et al. *End-to-end anti-spoofing with rawnet2*. in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. IEEE.

[69] Xie, Y., Z. Zhang, and Y. Yang. *Siamese Network with wav2vec Feature for Spoofing Speech Detection*. in *Interspeech*. 2021.

[70] Wang, C., et al. *Fully Automated End-to-End Fake Audio Detection*. in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 2022.

Rabbia Mahum received the B.Sc. degree in computer science from COMASTS University Wah in 2015. She achieved the MS degree in computer science from UET, Taxila in 2018. Besides this, she is pursuing PhD from UET, Taxila. Her research areas include Computer Vision, Deep Fake Audio Synthesis and Detection, and Medical Imaging.

Dr. Aun Irtaza has done Postdoctoral from the Department of Computer Science and Engineering, University of Michigan. He completed PhD from Fast University Islamabad. He is currently serving as Professor Associate in Computer Science Department, UET, Taxila. He has numerous research articles related to Computer Vision.

Dr. Ali Javed has received the B.Sc. degree (Hons.) in software engineering and the M.S. and Ph.D. degrees in computer engineering from the UET Taxila, Pakistan, in 2007, 2010, and 2016, respectively.,He is currently working as an Associate Professor with the Software Engineering Department, UET Taxila. He worked as the HOD with the Software Engineering Department, UET Taxila, in 2014. He worked as a Postdoctoral Scholar with the SMILES Laboratory, Oakland University, USA, in 2019.