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.2017.DOI

Deepfake Audio Detection via MFCC features using Machine Learning

AMEER HAMZA[1], ABDUL REHMAN JAVED[2*,3], FARKHUD IQBAL[4], NATALIA KRYVINSKA[5], AHMAD S. ALMADHOR[6], ZUNERA JALIL[2], ROUBA BORGHOL[7]

[1]Faculty of Computing and AI, Air University, Islamabad, Pakistan
[2]Department of Cyber Security, Air University, Islamabad, Pakistan
[3]Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon
[4]College of Technological Innovation, Zayed University, Abu Dhabi, UAE
[5]Information Systems Department, Faculty of Management, Comenius University in Bratislava, Odbojárov 10, 82005 Bratislava 25, Slovakia
[6]College of Computer and Information Sciences, Jouf University, Saudi Arabia
[7]Rochester Institute of Technology, Dubai Silicon Oasis, UAE

Corresponding author: abdulrehman.cs@au.edu.pk, natalia.kryvinska@fm.uniba.sk

ABSTRACT Deepfake content is created or altered synthetically using artificial intelligence (AI) approaches to appear real. It can include synthesizing audio, video, images, and text. Deepfakes may now produce natural-looking content, making them harder to identify. Much progress has been achieved in identifying video deepfakes in recent years; nevertheless, most investigations in detecting audio deepfakes have employed the ASVSpoof or AVSpoof dataset and various machine learning, deep learning, and deep learning algorithms. This research uses machine and deep learning-based approaches to identify deepfake audio. Mel-frequency cepstral coefficients (MFCCs) technique is used to acquire the most useful information from the audio. We choose the Fake-or-Real dataset, which is the most recent benchmark dataset. The dataset was created with a text-to-speech model and is divided into four sub-datasets: *for-rece*, *for-2-sec*, *for-norm* and *for-original*. These datasets are classified into sub-datasets mentioned above according to audio length and bit rate. The experimental results show that the support vector machine (SVM) outperformed the other machine learning (ML) models in terms of accuracy on *for-rece* and *for-2-sec* datasets, while the gradient boosting model performed very well using *for-norm* dataset. The VGG-16 model produced highly encouraging results when applied to the for-original dataset. The VGG-16 model outperforms other state-of-the-art approaches.

INDEX TERMS Deepfakes, Deepfake audio, Synthetic audio, Machine learning, Acoustic Data

I. INTRODUCTION

Deepfake is a portmanteau of *deep learning* and *fake*. Deepfake is a type of digitally-created content in which the original human faces in a photo, video, or recording have been swapped out for computer-generated ones [1], [2]. Deepfake first surfaced on Reddit in 2017 when a user named "deepfakes" submitted a falsified video on this website with a different actor's face. As a new technology, it unavoidably carries a slew of legal difficulties that infringe on personal interests like portraiture rights, reputation rights, and copyright and inflicts economic and reputational harm to businesses [3], [4]. Furthermore, a fabricated video of a politician or government being released will cause a media crisis, social instability, and national instability [5], [6], [7].

Audio deepfakes are AI-generated or modified audio that appears to be real. Since audio deepfakes have been employed in several criminal activities in recent years, the ability to detect them is crucial. Detecting deepfake in audio, video, and text is a broad and active research domain. Between 2018 and 2019, there was a significant increase in the number of articles about deepfake (from 60 to 309) [8]. Articles about deepfakes were expected to increase to over 730 by 2020's end, according to predictions made on July 24 [9]. In [10] found that most focus is on video deepfakes, particularly in developing video deepfakes.

Deep Fakes are increasingly detrimental to privacy, social security, and Authenticity. However, recent works have focused on deepfake video detection, achieving greater accuracy. However, audio spoofing and calls from malicious sources are generated through deep fakes, which need a specially trained model for handling this. The deepfake audio detection based purely on audio is less explored than image

and video-based approaches, as these works simultaneously utilize the audio and Spatio-temporal information in the video to train the deep learning model. However, only the audio-based classifier's classification and detection are very significant. Hence, to this end, we proposed an approach based on multiple machine learning algorithms to improve the accuracy of the classification models using Random Forest, Decision Tree, and SVM algorithms. We provide comparative results and analysis of the baseline models. We conducted our experiments on Fake-or-Real Dataset, and there were four sub-detests.

The ASVspoof2015 [11] is the first automatic speaker verification spoofing and countermeasures dataset that stimulates research in this field. It decreases the equal error rate (EER) by less the 1.5%. Some attacks have even 50% EER. However, unknown attacks can have five times more EER. Further, in ASVspoof2017 [12], the limits of replay spoofing attack detection are worked upon. The EER of 6.73% and the Instantaneous frequency cosine coefficients (EFCC) drastically improve countermeasures performance. Then ASVspoof2019 [13] put more emphasis on countermeasures concerning automatic speaker verification and spoofed audio detection. Other than that, computer vision algorithms such as convolutional neural networks (CNN) is used low-quality audio spectrograms for synthetic speech detection [14]. The time information can be lost in CNN-based models. Hence, probabilistic forecasting with a temporal convolutional neural network is used for improving automatic speaker verification and spoofed audio detection [15].

This research aims to derive a methodology for identifying deepfake audio from non-synthetic or real audio. It provides the following contributions to identify deepfake audios effectively by resolving the restrictions discussed above:

- Propose a transfer learning-based approach to detect deepfake.
- Extend work on deepfake audio detection on the Fake-or-Real dataset by conducting detailed experiments on Fake-or-Real datasets and sub-datasets using machine and deep learning-based approaches.
- Use a superior feature extraction approach to obtain MFCC features from audio sources.
- Results reveal that the SVM model outperforms other ML models compared to other dataset sub-sets except for the for-original dataset. The VGG-16 model produced highly encouraging results when applied to the for-original dataset.

The paper will proceed as described below. Section II explains the literature review methods. The suggested approach and algorithms are described in III. The analysis and results of the experiments are provided in Section IV. Section V presents the discussion on the proposed approach. Finally, Section VI presents the overall conclusion.

II. LITERATURE REVIEW

Audio deepfakes audio is generated, edited, or synthesized using artificial intelligence, which appears real. Detecting audio deepfakes is critical since audio deepfakes have been used in several illegal actions in banking, customer service, and call centers. To detect audio deepfakes, one must first understand the procedures of generation. As the name suggests, audio deepfake algorithms are classified into three types: Replay attack, speech synthesis, and voice conversion are all possible. This section gives the reader each subcategory's most recent and relevant frameworks.

Audio forensics is a branch of forensics used to authenticate, enhance, and analyze audio information to aid in investigating various crimes. Audio as forensic evidence must be modified and analyzed before criminal prosecution. However, more significantly, it must be validated to demonstrate that it is genuine and has not been tampered with. Several methods, primarily employing AI/ML-based techniques, have been used to detect audio events in the last decade. A deep learning framework was employed by the authors of the study [16] for audio-deep fake detection. The model separability is increased using a Long-short term memory (LSTM)-the based network is used to recognize events in sub-sampled signals [17]. To reduce the audio signal complexity and ease of reconstruction encode, the frequencies higher than the Nyquist frequency[18] are used, and the authors [19] utilized non-uniform sampling for audio subsampling.

Replay attacks consist of repeatedly playing back a recording of the voice of the intended victim. Replay attacks come in two forms, the first is far field detection, and the second is copy-and-paste detection [20], [21]. As of now, deep convolutional networks are used as a method for detecting complete replay attacks [22]. Several methods have been developed for identifying replay attacks, and they center on the characteristics that are provided in the network. The method of using deep convolutional networks to detect replay attacks was found to have an Equal Error Rate (EER) of zero percent on the ASVspoof2017 training and test dataset [12].

Speech synthesis (SS) is recreating human speech digitally, typically using computer software or hardware. TTS is a component of SS that takes in written material and outputs spoken language based on that text according to predetermined linguistic rules. Text reading and AI personal assistants are just two applications of speech synthesis. Another perk of speech synthesis is that it can mimic various voices and dialects without relying on canned recordings. Lyrebird [1], a powerful speech synthesis company, employs deep learning models to synthesize 1,000 sentences in a second. TTS largely relies on the quality of the speech corpus to build the system, and regrettably, creating speech corpora is expensive [23]. Speech synthesis is the artificial reproduction of human speech using software or hardware system programs. In order to synthesize 1,000 sentences per

[1]https://www.descript.com/lyrebird

second, Lyrebird uses deep learning models. The success of a TTS system is highly dependent on the quality of the speech corpus upon which it is built, and it is costly to collect and annotate speech samples. Char2Wav is a framework for speech synthesis production from start to finish. PixleCNN is also the foundation of WaveNet [24], an SS framework. WaveGlow prioritizes stage two of the two-stage process generally used by text-to-speech synthesis systems (encoder and decoder). Therefore, WaveGlow is concerned with modifying specific time-aligned data. Incorporating information into sound files by using encoding techniques like a mel-spectrogram. The Tacotron 2 [25] system comprises two parts. The first component is an attention-based recurrent sequence-to-sequence feature prediction network. This component's output is a mel anticipated sequence. Frames of a spectrogram A modified WaveNet vocoder is the second component. For audio data, [26], [27] used GAN-based generative models. It operates on Mel spectrograms and employs a fully convolutional feed-forward network as the generator. The authors give a summary of their recently created data set. It comprises 117,985 created audio segments of 16-bit Pulse Code Modulation(PCM) wav format and is available on zenodo [2].

The current study has poor performance validation and testing results detecting deep false audios. Feature-based techniques are required to improve the outputs of machine learning models. The deep learning approaches show better results but require greater training time and computational resources. Hence, the potential for machine learning approaches in deepfake detection is explored, while the limitation of handling higher feature sets and complexities can be solved through a transfer learning-based deep learning approach.

III. PROPOSED METHODOLOGY

In machine learning, training a model always involves the trade-off of over-fitting and under-fitting, which negatively impacts the model's real-time performance. It is difficult to handle this trade-off so that models do not over-fit or under-fit. One of the major issues in deepfake is the high false-positive rate ratio, which occurs when most models classify an unseen pattern as abnormal if it is not included in the training set. It is due to the model's inability to be trained on a large dataset. A dataset that covers all possible patterns and cases, deepfake and real. It is regarded as a theoretical concept that cannot be implemented practically. Hence, the dataset Fake-or-Real [28] is divided into four datasets:*for-rece*, *for-2-sec*, *for-norm*, *for-original*, where *for-original* dataset is the collection of other three datasets and without much preprocessing.

This research aimed to develop a technique to classify deep fake synthetic audio under different background noise and audio sizes and duration. We proposed a framework that handles the big data training set and performs detection using different supervised and unsupervised machine learning algorithms. The following section explains the proposed framework for all sub-datasets, including data handling, preprocessing, feature engineering, and the classification phase. Figure 1 Shows the detailed architecture diagram of the proposed framework, consisting of 1) data preprocessing, 2) feature extraction 3) Classification models. The detailed description of each phase is as follows:

A. DATA PREPROCESSING

More than 195,000 real human and synthetic computer-generated speech samples are included in the Fake-or-Real (FoR) collection. Classifiers may be trained on the dataset to identify fake speech better. Information from Deep Voice 3 is included [29] and Google Wavenet[24] TTS and various human sound recordings. This dataset may be accessed in four different varieties[1) for-original,2) for-norm, 3) for-2sec, and 4) for-rerec]. The original version includes the files without any changes from when they were first extracted from the speech sources. The latest volume (For-norm) contains the duplicate files as the first, but they have been standardized in terms of sampling rate, volume, and various channels to achieve gender and class parity. The second is the basis for the third (for-2sec), except that the files are truncated after 2 seconds instead of the original length. The third and final version (for-rerec) is a re-recorded version of the for-2second dataset created to simulate an attacker transmitting an utterance via a voice channel. However, these datasets suffer from duplicate files, 0-bit files, and different bit-rate in audio signals. They negatively affect the ML model's training and performance. Hence, we preprocess the dataset to remove the duplicate and 0-bit file, which does not contribute to model training. Also, the bit rate is standardized to zero-padding for an audio waveform with less than 16,000 samples, conforming to an operationally viable bit rate for the TensorFlow audio signal processing library. Also, the data is normalized using a standard scaler to ease model training.

B. FEATURE EXTRACTION

Deepfake audio signal often consists of similar feature sets to the original signal. However, distinguishability is challenging to advance in deep learning approaches in generating deepfakes. Hence, extracted features can strongly affect the model's predictive power and accuracy. It is observed that audio signals in the frequency domain can provide us the features which are helpful in the detection and classification of deepfake audios, which can deceive a human under specific scenarios. For this purpose, we use Mel-frequency Cepstral Coefficient (MFCC), a widely used feature for speech recognition [30], [31]. The dataset (Fake or Real Audio dataset) used in this study is a more recent dataset, which was only used once in research. This study is not limited to only MFCC features; we also employed cepstral (MFCC), Spectral (Roll-off point, centroid, contrast, bandwidth), Raw signal (zero cross rate), and signal energy features and made a featured ensemble, but our primary focus is on MFCC

[2]https://zenodo.org/record/5642694

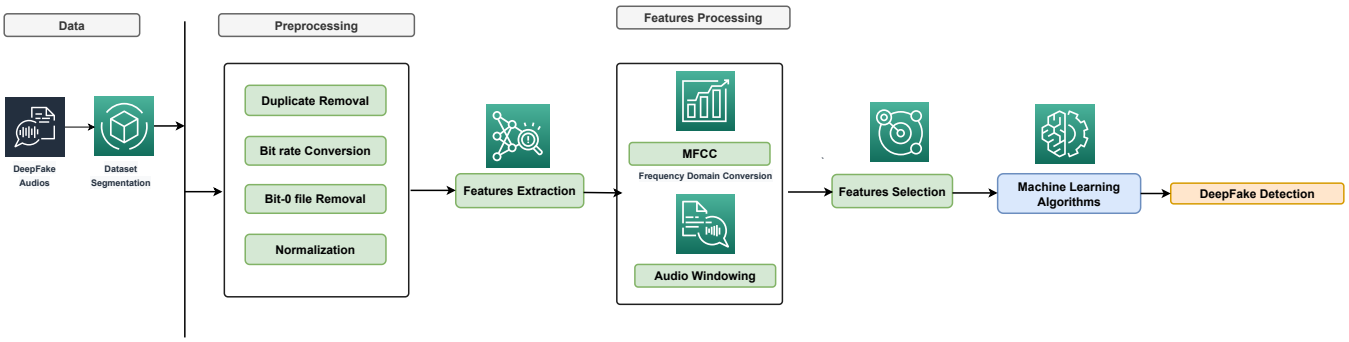


FIGURE 1: Graphical Representation of Proposed Approach for detection of deepfake audios

features because MFCC used log function and Mel-filter to mimic the human hearing system furthermore MFCC apply triangular band-pass filters to covert the frequency information to mimic what a human perceived. Figure 2 shows the MFCC series of audio files; also, for showing the auditory power of the signal, the amplitudes are presented in decibels (db). In this work, MFCC is used for deepfake audio detection. Following the initial processing of audio signals, a vector group representing the MFCC will be generated out of each frame of the sound waveform. This study uses Mel-frequency cepstral coefficient and short-time Fourier transform (STFT), which transform the waveforms from the time-domain signals into the time-frequency-domain signals. In Figure 3, the comparison between the fake and accurate audio signals in terms of their spectrogram representation is shown. First, the spectrogram is shown in terms of their different amplitude, and then, for better auditory inspection, the signal is further analyzed using the decibel (db) of the given signal. It helps us understand which auditory features are relevant in distinguishing between the deepfake and real audio signal. In this section, we explain the feature extraction and selection process. In our case, the sampling (frame) rate is 44100. The 270 retrieved features of each audio file are stored in a data frame. We reduced the characteristics to only those that would be beneficial and got rid of the rest using Principal Component Analysis (PCA) [32]. 65 characteristics are crucial enough to be sent to deepfake detection models. We employ PCA's explained variance ratio metric to determine the value of carefully chosen features. The value of explained variance ratio is (97%), indicating that the selected data's usefulness is convincing. We set different values of PCA and find the most important features where n_component is 65.

C. CLASSIFICATION MODELS

1) Random Forest

Random Forest is a decision tree-based algorithm except that it fits many categorizing decision trees on different sub-samples of the dataset and then uses averaging to integrate all the decision trees. It helps in the mitigation of dataset overfitting problems. Random forest is used in calculating

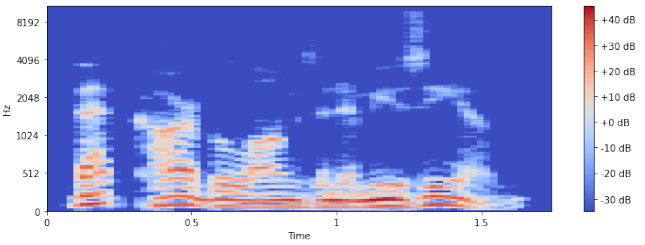


FIGURE 2: The melspectrogram representation of audio signal where the amplitude is depicted in terms of decibel

the feature importance by adding the gain of each feature and scaling the number of samples passing through the node. Let us assume that k is the node, X_j is the importance of features, and the total samples toward all nodes are Y_k. The importance of a feature can be represented as in equation 1:

$$X_j = \sum k : jY_k G_K \tag{1}$$

where, nodes k split on feature j. The final feature importance X_j for each feature is calculated by normalizing the X'_j for each tree and then summing those normalized values for each tree in the random forest.

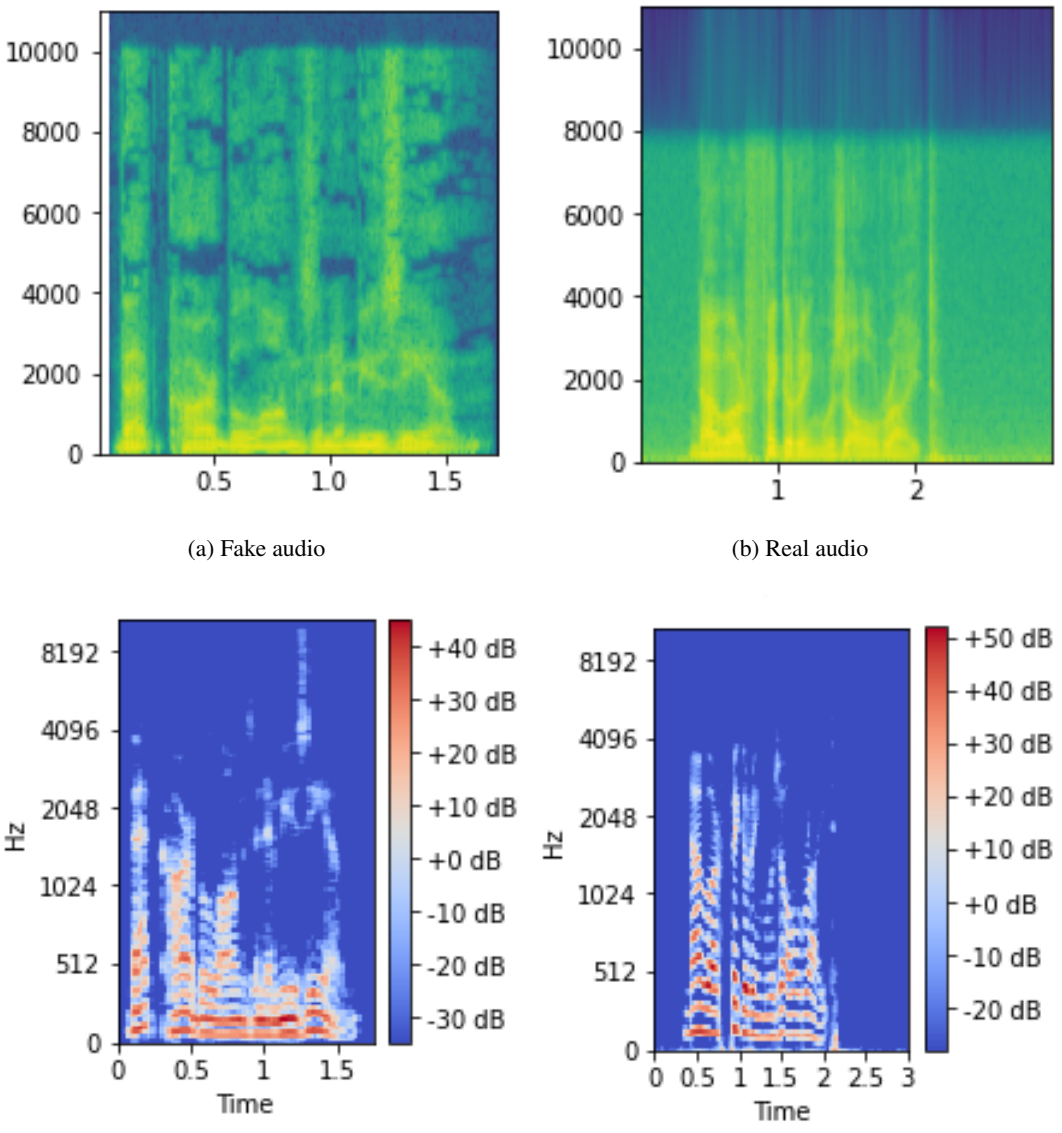
$$X_{j'} = \frac{X_{j'}}{\sum_z X_{j_z}} \tag{2}$$

$$RFX_{jj'} = \frac{\sum_z X j'_{jz}}{\sum_{z,t} X j'_{jzt}} \tag{3}$$

As in equation 2 and 3, z indicates all features, and t depicts all trees in a random forest. The X_j is the importance of a feature for node j, while $X_{j'}$ is its normalized feature importance. $RFX_{jj'}$ is the feature importance for all trees in a random forest. Moreover, X_{j_z} is normalized importance of feature j w.r.t tree t. The model makes predictions based on the important features obtained, as mentioned in Figure 3.

2) Support Vector Machine (SVM)

SVM is a supervised learning method that relies primarily on two assumptions: 1) Converting data into a high-dimensional space may reduce complex classification issues with complex



(a) Fake audio

(b) Real audio

(c) Fake Audio

(d) Real Audio

FIGURE 3: In (a) and (b), the comparison is shown between the deepfake and real audio signal in spectrogram where the difference in amplitude is apparent. In (c) and (d), the amplitude is shown in terms of decibels (db) for understanding the auditory parts of the audio signal.

decision surfaces to more minor problems that may be solved by making it linearly separable, and 2) only training patterns near the decision surface provide the most sensitive details for classification. Assume a deepfake detection problem as a binary classification with linearly separable vectors $x_i \in \mathbb{R}^n$, as the decision surface used to classify a pattern as belonging to one of the two classes is the hyperplane H_0. If x is a random vector $n * \mathbb{R}$, we define

$$f_{(x)} = w.x + b \tag{4}$$

Dot product is represented d $(.)$ in equation 4. The set of all x-vectors that satisfy the equation $f_{(x)} = 0$ is denoted by H_0. Assuming two hyperplanes, H_1 and H_2, the distance between them is referred to as their *margin* which can be represented as follows:

$$\frac{2}{\|w\|} \tag{5}$$

The decision hyperplane H_0 depends on vectors closest to the two parallel hyperplanes called support vectors. The margin must be maximal to obtain a classifier that is not much adapted to the training data. Consider a collection of training data vectors $X = x_i, ...x_L, x_i \in \mathbb{R}^n$ and a set of

matching labels, $Y = y_i, ...y_L, y_i \in \{1, -1\}$. We consider the hyperplane H_0 to be optimally separated if the vectors are categorized without error and the margin is greatest. The vectors must verify in order to be accurately categorized.

$$f_{x_i} > +1 \; for \; y_i = +1 \tag{6}$$

$$f_{x_i} > -1 \; for \; y_i = -1 \tag{7}$$

Hence, finding the SVM classifying function H_0 can be stated as follows:

$$minimize \; \frac{1}{2} \|w\|^2 \tag{8}$$

$$y_i f_{(x_i)} \geq 1, \; \forall_i \tag{9}$$

The SVM was chosen for its properties that aid in classifying deepfake audios. It performs well with a clear margin of separation between samples and is effective in high-dimensional environments. It employs a subset of training points in the decision function, making it memory efficient. It works well when the number of dimensions is more than the size of the sample set. SVM does not perform very well on our *for-original* dataset data set because the required training time and the noise in the data set are higher. It does not directly provide probability estimates, calculated using an expensive five-fold cross-validation that takes a long time to train. However, the clean datasets extracted from *for-original* dataset perform better on the classification task. SVM has been shown to perform effectively in higher-dimensional data, most notably when detecting events in audio data. Hence, for deepfake audio, we implemented it by utilizing the Scikit-learn library. We use radial basis function (RBF) kernel, C=4, and probability=True.

3) Multi-layer Perceptron (MLP)

MLP is adequate for classification tasks; a multilayer perceptron, through layers, can effectively filter the relevant features from data and tune the parameters of the models for optimal predictions. There are at least three levels in the MLP model: an input layer, a hidden layer of calculation nodes, and an output layer of processing nodes. In this study, we use hyperparameters of MLP classifier as a hidden-layer-size staple:length= 100, solver=adam and RMSprop, while RMSprop is used for smaller datasets, shuffle=True and verbose=False, activation-function=relu.

4) Extreme Gradient Boosting (XGB)

XGB is a parallel and optimized version of gradient boosting algorithms that combines efficiency and resource management. It implements gradient-boosted decision trees in an iterative model by combining weak base models into a stronger learner. The residual is utilized to refine the loss function and improve the prior prediction at each iteration of the gradient boosting algorithm. We use a learning rate of 0.1 and an estimator of 10000 for the XGBoost algorithm. However, it is vulnerable to outliers because each successive classifier is compelled to correct the mistakes made by its prior learners. This is because the estimators rely on historical predictions to determine their accuracy. For this reason, streamlining the process is complex.

IV. EXPERIMENTS AND RESULTS

About 195,000 human and synthetic speech samples were used to create the Fake-or-Real (FoR) dataset. In Table 1, we offer a summary of the data set. Classifiers may be trained on the dataset to identify fake speech better. The dataset is an amalgamation of information from the following recent datasets: first, Text-to-speech programs, such as Deep Voice 3 and Google Wavenet TTS [29]. Secondly, includes many different types of the recorded human voice, including those from the Arctic Dataset, LJSpeech Dataset, VoxForge Dataset, and user-submitted recordings [33], [34], [35]. The four dataset versions available for public consumption are for-original, for-norm, for-2sec, and for-rerec. The for-original folder stores the raw data from the speech sources.

The for-norm has some duplicate files but is otherwise well-balanced across demographic categories (gender and socioeconomic status) and technical parameters (sample rate, volume, and multiple channels). The third one is like the second one, only the files are cut off after 2 seconds, and it is called for-2sec. The last variant, dubbed for-here, is a re-recording of the for-2second dataset meant to mimic a situation in which an attacker transmits a speech over a vocal channel like a phone call or voice message. We provide the outcomes of our binary classification analysis of the suggested method. Table 2 shows the experimental findings for spotting deepfakes.

The experiments were also performed using noisy audio sound signals. For this purpose, we added synthetic noise to each audio signal of three datasets (for-2sec, for-norm, and for-rerec dataset). This method kept both original and noisy audio in the dataset and increased the audio signal sample. The length of the original for-2sec dataset is 17870 audio samples, and after adding noise to the dataset, the new dataset will be composed of 35740 audio samples, the same for the for-rerec and for-norm datasets.

A. FOR-REREC DATASET

The results of the for-rerec dataset are presented in Table 2. Multiple ML models are applied to obtain better results. The machine learning algorithms such as Support Vector Machine (SVM) have 98.83% accuracy, Decision Tree 88.28%, Random Forest Classifier 96.60%, AdaBoot 87.67%, Gradient Booting 93.51%, XGB Classifier 93.40%. The SVM model exhibited the highest results using the for-rerec dataset.

The result of the for-rerec dataset noisy audio signals classification is presented in Table 3. Results depict that the MLP and SVM models obtained the highest accuracy score of 98.66% and 98.43% compared to other ML models. The other ML models like; DT, LR, and XGB obtained 82.12%, 88%, and 88.92% accuracy.

TABLE 1: Dataset Description

Datasets	Size	Description
FOR-REREC DATASET	1.5 GB	It is a re-recorded version of the for-2-second dataset to simulate a scenario where an attacker sends an utterance through a voice channel (i.e., a phone call or a voice message).
FOR-2SEC DATASET	1 GB	Contains audios based on FOR-NORM dataset, but with the files truncated at 2 seconds
FOR-NORM DATASET	5.8 GB	the same files as FOR-ORIGINAL dataset, but balanced according to gender and the same sampling rate, volume, and channels for each.
FOR-ORIGINAL DATASET	7.7 GB	The audios are collected from various sources, without any modification

B. FOR-2SEC DATASET

In for-2sec dataset consist of audio with two-second intervals. The audio is complex, as the information in that small interval is little. However, it is much easier for machine learning algorithms to process data in this form. Hence, we observe better performance. The results are depicted in Table 2. We observe MLP classifier accuracy of 94.69, Random Forest of 94.44, and SVM 97.57. gradient boosting of 94.30 and Adaboosing of 90.23. The MLP model outperforms the other ML model in terms of accuracy.

Table 3 shows the results of for-2sec dataset with noisy audio signal classification. To get better outcomes, several ML models are used. The ML algorithms such as SVM obtained 99.59% accuracy, MLP obtained 99.49%, DT 87.52% accuracy, and so on. The SVM exhibited the highest accuracy compared to other ML models using noisy audio signals.

C. FOR-NORM DATASET

It contains recorded audio at 12-second intervals. The result of the for-norm dataset is shown in Table 2. It shows MLP Classifier 86.82, Random Forest Classifier 90.60, extra trees 91.46, Gradient Booting 92.63, XGB Classifier 92.60, LDA 91.35, Gaussian NB 81.81, and Adabost 89.40. However, some algorithms show average results, like QDA 61.36 and KNN 64.21. The Gradient Boosting classifier obtained the highest results compared to the other ML models.

The results of noisy audio from the for-norm dataset are presented in Table 3. The results of the for-norm dataset are less than the other two datasets. The XGB model obtained the highest results using noisy audio from the for-norm dataset. All other ML models obtained quite well results but not so impressive.

TABLE 2: Accuracy comparison for machine learning models

Models	for-2sec	for-norm	for-rerec
SVM	97.57	71.54	98.83
MLP Classifier	94.69	86.82	98.79
Decision Tree	87.13	62.16	88.28
Extra Tree Classifier	94.61	91.46	96.87
Gaussian Naive Bayes	88.20	81.81	81.91
Ada Boost	90.23	88.40	87.67
Gradient Boosting	94.30	92.63	93.51
XGBoost	94.52	92.60	93.40
Linear Discriminant Analysis	89.50	91.35	87.56
Quadratic Discriminant Analysis	96.13	61.36	96.91

TABLE 3: Accuracy comparison for noisy audio signals using machine learning models

Models	for-2sec	for-norm	for-rerec
SVM	99.59	75.21	98.43
MLP Classifier	99.49	89.22	98.66
Decision Tree	87.52	65.10	82.12
Extra Tree Classifier	97.91	90.19	96.25
Logistic Regression	87.53	86.28	88.00
Gaussian Naive Bayes	79.77	80.16	82.14
Ada Boost	85.28	91.35	83.89
Gradient Boosting	92.29	93.50	88.86
XGBoost	92.22	94.25	88.92
Linear Discriminant Analysis	87.05	90.52	85.88
Quadratic Discriminant Analysis	96.22	65.15	95.59

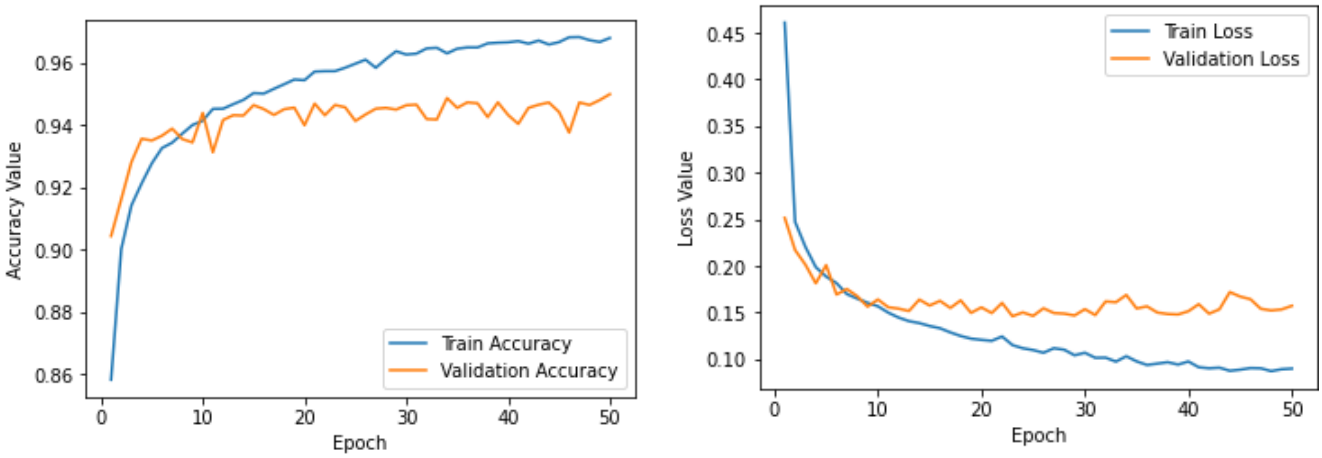
D. FOR-ORIGINAL DATASET

The *for-original* datasets are compiled from various datasets and consist of audio samples of various lengths, bit-rates, and noise levels. The machine learning models did not produce comparatively better results. These models would not be able to handle this data's complexities and feature variations. To this end, We used a transfer learning-based deep learning approach. In this approach, we extracted the same visual features of MFCC from the audio data. These visual features train the VGG-16-based model and LSTM to perform deepfake or real audio classification. Finally, the VGG16 model outperformed the LSTM model with a testing accuracy of 93%. The LSTM model obtained 91% accuracy. The VGG-16 model uses ImageNet weights and input shapes (64 x 64 x 3). The validation accuracy of 0.94 and validation loss of 0.14 is obtained, while the testing accuracy is 93%. Figure 4a shows the training and validation graph, while Figure 4b shows the training and validation loss of the VGG16 model.

E. MODEL COMPARISON

This study compares the model accuracy with the other baseline paper [36] to assess the efficacy of our proposed model. It is easier to compare results when the experimental conditions (dataset, data samples) are identical to those used in the initial study. As presented in this section, the dataset utilized in this investigation has only been used once in a previous study [36]. Because of this reason, the suggested method cannot be compared to any other studies.

Our technique shows potential in terms of classification accuracy. This work obtained comparatively better results in ensemble-based machine learning models such as boosting algorithms, as the XGboost algorithm shows greater accuracy than the baseline model. The model's accuracies for three sub-datasets are shown in Table 2 and 3.



(a) Model accuracy

(b) Model loss

FIGURE 4: Comparison between the validation and training (accuracy and loss)

Table 2 and 3 compare the machine learning model results of the feature-based approach to training machine learning algorithms. Our approach to selecting the best feature and ML classifiers obtained promising results on three datasets (FOR-REREC DATASET, OR-2SEC DATASET, and FOR-NORM DATASET). However, The *for-norm* dataset does not perform well on our approach using a simple SVM algorithm as the data is of high dimensions. Without the dimensionality reduction on a complex dataset, it performs poorly. This dataset contains audio of a length greater than 12 seconds. Hence, the windowing technique can perform better in combination with MFCC. The proposed approach is compared with the baseline approach that used FOR-ORIGINAL DATASET for experimentation [36]. The existing approach used various ML models (SVM, RF, KNN, XGB) to detect deepfake from FOR-ORIGINAL-DATASET. The proposed approach obtains the highest testing score of 93%, which is 26% higher than the best score of existing work using the SVM model. It is concluded that the proposed approach can efficiently detect deepfake audio. The dataset used in this study is only used in only one previous study. The proposed and existing approaches' experimental settings are similar (dataset, data split). In addition, the comparative analysis of the proposed method with the state-of-the-art feature extraction techniques is presented in Table 4. The proposed approach combines features from multiple feature extraction techniques and extracts the most optimal features for classification. The two deep learning models are employed in this research. The proposed approach employed VGG16 and LSTM model with a feature ensemble of MFCC-40, Roll-off point, centroid, contrast, and bandwidth features. The features extracted from each method are combined for model classification. The VGG16 model obtained the highest results compared to the existing study with an accuracy of 93%. Furthermore, the LSTM model obtained an accuracy of 91%. The existing approach proposed by khochare et al. (2021) used MFCC features and various machine learning models for deepfake audio detection [36]. They have utilized 20 MFCC features for each audio. The author employed multiple machine learning models (SVM, RF, KNN, and XGB). The author used 20 MFCC features with the SVM model and obtained the highest accuracy rate of 67%. Another study proposed by Reimao et al. (2019) used both machine learning and deep learning techniques along with various feature extraction methods [28]. The author used Timbre Model Analysis (Brightness, Hardness, Depth, Roughness) features with multiple ML models (NB, SVM, DT, and RF). According to the ML model classification results, the SVM model using the various feature extraction methods obtained a 73.46% accuracy rate. Furthermore, STFT, Mel-Spectrograms, MFCC, and CQT feature extraction methods are used with the VGG19 model and obtained 89.79% accuracy. Compared to the previous research, the VGG16 model achieved the highest results, with an accuracy of 93%. More so, the LSTM model achieved 91% accuracy. The VGG16 model loss and training and validation accuracy are shown in Figure 4. The proposed approach with the features mentioned in section III-B outperforms the previous state-of-the-art feature extraction techniques is presented in Table 4.

V. DISCUSSION

This research extended the work on deepfake audios by extending the work on the Fake-or-Real dataset. This dataset comprises a state-of-the-art dataset in audio detection and classification. We improved upon the algorithm's performance, which was previously trained on feature-based approaches by using the MFCC-based features, indicating considerable improvements in inaccuracy. Our feature outperforms the feature-based approach by 10 to 20 percent on average across these datasets. The for-norm dataset performs poorly on our approach using simple SVM algorithms. Win-

TABLE 4: Comparison between results of the proposed approach and existing approach

Approaches	Features	Models	Accuracy (%)
Existing Approach [36]	MFCC-20	SVM	67
		RF	62
		KNN	62
		XGB	59
Existing Approach [28]	Timbre Model Analysis (Brightness, Hardness, Depth, Roughness)	NB	67.27
		SVM	73.46
		DT (J48)	70.26
		RF	71.47
	STFT, Mel-Spectrograms, MFCC and CQT	VGG19	89.79
Proposed Approach	MFCC-40, Roll-off point, centroid, contrast, bandwidth	LSTM	91
		VGG16	93

dowing techniques, in combination with MFCC, can perform better. We conduct additional experiments on machine learning algorithms categorizing into (1) Statistical models like QDA, LDA, and Gaussian Naive Bayes for dimensionality reduction to reduce noise in the data. Then (2) Tree-based models such as Decision Tree, Extra Tree, and Random Forest these algorithms can handle multidimensional data. Therefore, they do not involve domain knowledge or parameter setting and are appropriate for exploratory pattern detection. Lastly, (3) Boosting models, namely Ada Boost, Gradient Boosting, and XGBoost, these algorithms fundamentally create several weak learners and combine their predictions with building a strong rule, which helps increase the accuracy of a model on feature-rich audio data. These three classes of ML algorithms are chosen for our approach to explore and improve these performances on MFCC-based feature sets. Besides this, we proposed a VGG-16-based deep learning model for the bigger dataset, which is the superset of the other three datasets. It uses transfer learning and trained on MFCC images feature for training the model. We obtained an accuracy of 93% while using half of the original dataset. A large amount of data correlated with higher model accuracy. We tried to obtain a limited performance dataset. The entire dataset can be explored for even better results in the future.

VI. CONCLUSION

The detection of audio data is significant as an essential tool for enhancing security against scamming and spoofing. Deepfake audios have garnered significant public attention as society rapidly recognizes its possible security danger. However, deepfake audio is extensively studied in combination with Spatio-temporal data of video. This study improves upon the Fake-or-Real (FoR) dataset, which comprises state-of-the audio datasets and custom audios for deepfake audio classification. It is further compiled into four sub-datasets. This study conducted experiments with multiple audio data features to detect deepfakes in audio data. This work extracts MFCC features from audio for feature engineering. Several machine learning algorithms are applied to the selected feature set to detect the deepfake audio. This approach gave higher accuracy and results in all cases than other state-of-the-art studies for audio data. This study obtained 97.57% accuracy with SVM using the for-2sec dataset compared to other ML models, while 92.63% was obtained by the Gradient Boosting classifier using the for-norm dataset. This study obtained 98.83% highest accuracy using the SVM model on the for-rerec dataset. We plan to explore the different window sizes for MFCC and various input sizes for models in the future. Future work can be done on evaluating these models against potential fluctuation and distortion in the audio signal, understanding which signal is greater. Moreover, studies on the state-of-the-art few-shot learning and Bidirectional Encoder Representations from Transformers (BERT) based models can be conducted. Furthermore, we plan to evaluate our models in ambient noise and reverberation circumstances. We intend to use feature extraction methods like i-vector, x-vector, a combination of MFCC and GFCC, and a combination of DWT and MFCC, which were not taken into account in the current scenario of experiments because it is the beginning of our journey to identify Deepfake audio.

REFERENCES

[1] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," IEEE Access, vol. 10, pp. 38885–38894, 2022.

[2] A. R. Javed, W. Ahmed, M. Alazab, Z. Jalil, K. Kifayat, and T. R. Gadekallu, "A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions," IEEE Access, 2022.

[3] A. R. Javed, Z. Jalil, W. Zehra, T. R. Gadekallu, D. Y. Suh, and M. J. Piran, "A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions," Engineering Applications of Artificial Intelligence, vol. 106, p. 104456, 2021.

[4] A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, "Privacy of web browsers: a challenge in digital forensics," in International Conference on Genetic and Evolutionary Computing, pp. 493–504, Springer, 2021.

[5] A. R. Javed, F. Shahzad, S. ur Rehman, Y. B. Zikria, I. Razzak, Z. Jalil, and G. Xu, "Future smart cities requirements, emerging technologies, applications, challenges, and future aspects," Cities, vol. 129, p. 103794, 2022.

[6] A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," Scientific Reports, vol. 12, no. 1, pp. 1–16, 2022.

[7] S. Anwar, M. O. Beg, K. Saleem, Z. Ahmed, A. R. Javed, and U. Tariq, "Social relationship analysis using state-of-the-art embeddings," Transactions on Asian and Low-Resource Language Information Processing, 2022.

[8] C. Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," The Wall Street Journal, vol. 30, no. 08, 2019.

[9] T. T. Nguyen, Q. V. H. Nguyen, C. M. Nguyen, D. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection: A survey," arXiv preprint arXiv:1909.11573, 2019.

[10] Z. Khanjani, G. Watson, and V. P. Janeja, "How deep are the fakes? focusing on audio deepfake: A survey," arXiv preprint arXiv:2111.14203, 2021.
[11] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in Sixteenth annual conference of the international speech communication association, 2015.
[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," ISCA (the International Speech Communication Association), 2017.
[13] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," tech. rep., Tech. Rep., 2019.[Online]. Available: http://http://www. asvspoof. org . . . , 2019.
[14] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," IEEE Signal Processing Letters, vol. 26, no. 1, pp. 94–98, 2018.
[15] Y. Chen, Y. Kang, Y. Chen, and Z. Wang, "Probabilistic forecasting with temporal convolutional neural network," Neurocomputing, vol. 399, pp. 491–501, 2020.
[16] Y. Kawaguchi, "Anomaly detection based on feature reconstruction from subsampled audio signals," in 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2524–2528, IEEE, 2018.
[17] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?," in 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6, IEEE, 2017.
[18] H. Landau, "Sampling, data transmission, and the nyquist rate," Proceedings of the IEEE, vol. 55, no. 10, pp. 1701–1706, 1967.
[19] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using dnn classifiers and dynamic acoustic features," IEEE transactions on neural networks and learning systems, vol. 29, no. 10, pp. 4633–4644, 2017.
[20] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 3, pp. 1–26, 2019.
[21] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in 2011 Carnahan Conference on Security Technology, pp. 1–8, IEEE, 2011.
[22] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention.," in Interspeech, pp. 681–685, 2018.
[23] K. Kuligowska, P. Kisielewicz, and A. Włodarz, "Speech synthesis systems: disadvantages and limitations," Int J Res Eng Technol (UAE), vol. 7, pp. 234–239, 2018.
[24] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.
[25] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779–4783, IEEE, 2018.
[26] J. Frank and L. Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," arXiv preprint arXiv:2111.02813, 2021.
[27] M. Hassaballah, M. A. Hameed, and M. H. Alkinani, "Introduction to digital image steganography," in Digital Media Steganography, pp. 1–15, Elsevier, 2020.
[28] R. Reimao and V. Tzerpos, "For: A dataset for synthetic speech detection," in 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1–10, IEEE, 2019.
[29] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," arXiv preprint arXiv:1710.07654, 2017.
[30] F. M. Rammo and M. N. Al-Hamdani, "Detecting the speaker language using cnn deep learning algorithm," Iraqi Journal For Computer Science and Mathematics, vol. 3, no. 1, pp. 43–52, 2022.
[31] Z. A. Abbood, B. T. Yasen, M. R. Ahmed, A. D. Duru, et al., "Speaker identification model based on deep neural networks," Iraqi Journal For Computer Science and Mathematics, vol. 3, no. 1, pp. 108–114, 2022.
[32] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of mfcc feature extraction accuracy using pca in indonesian speech recognition," in 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 379–383, IEEE, 2018.
[33] J. Kominek and A. W. Black, "The cmu arctic speech databases," in Fifth ISCA workshop on speech synthesis, 2004.
[34] K. Ito and L. Johnson, "The lj speech dataset." https://keithito.com/LJ-Speech-Dataset/, 2017.
[35] K. MacLean, "Voxforge," Ken MacLean.[Online]. Available: http://www.voxforge.org/home.[Acedido em 2012], 2018.
[36] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A deep learning framework for audio deepfake detection," Arabian Journal for Science and Engineering, pp. 1–12, 2021.

AMEER HAMZA is with the Department of Creative Technology, Air University, Islamabad. He is doing his Master's degree in Artificial Intelligence from Air University, Islamabad, Pakistan.

ABDUL REHMAN JAVED is a lecturer at the Department of Cyber Security, Air University, Islamabad, Pakistan. He has worked with National Cybercrimes and Forensics Laboratory at Air University, Islamabad, Pakistan. He received his Master's degree in Computer Science from the National University of Computer and Emerging Sciences, Islamabad, Pakistan. He is a member of both IEEE and ACM. He is a cybersecurity researcher and practitioner with industry and academic experience. He has reviewed over 150 scientific research articles for various well-known journals. He is a TPC member of CID2021 (Fourth International Workshop on Cybercrime Investigation and Digital forensics - CID2021) and the 44th International Conference on Telecommunications and Signal Processing. He has served as moderator in the 1st IEEE International Conference on Cyber Warfare and Security (ICCWS). He has authored over 50 peer-reviewed research articles and is supervising/co-supervising several graduate (BS and MS) students on health informatics, cybersecurity, mobile computing, and digital forensics topics. His current research interests include but are not limited to mobile and ubiquitous computing, data analysis, knowledge discovery, data mining, natural language processing, smart homes, and their applications in human activity analysis, human motion analysis, and e-health. He aims to contribute to interdisciplinary research in computer science and human-related disciplines.

FARKHUND IQBAL holds the position of Associate Professor in the College of Technological Innovation, Zayed University, United Arab Emirates. He is an Affiliate Professor at the School of Information Studies, McGill University, Canada, and an Adjunct Professor at the Faculty of Business and IT, Ontario Tech University, Canada. He leads the Cybersecurity and Digital Forensics (CAD) research group at the Center for Smart Cities and Intelligent Systems, Zayed University. He holds a Master's (2005) and a Ph.D. degree (2011) from Concordia University, Canada. He uses Artificial Intelligence, Machine Learning, and Data Analytics techniques for problem-solving in cybersecurity, health care, and cybercrime investigation in the smart city domain. He has published more than 120 papers in high-ranked journals and conferences. He has served as a chair and co-chair for several IEEE/ACM conferences and has been a guest editor and reviewer for multiple high-rank journals.

NATALIA KRYVINSKA is a Full Professor and a Head of the Information Systems Department at the Faculty of Management, Comenius University in Bratislava, Slovakia. Previously, she served as a University Lecturer and a Senior Researcher at the eBusiness Department at the University of Vienna's School of Business Economics and Statistics. She received her Ph.D. in Electrical & IT Engineering from the Vienna University of Technology in Austria, and a Docent title (Habilitation) in Management Information Systems from the Comenius University in Bratislava, Slovakia. She got her Professor title and was appointed for the professorship by the President of the Slovak Republic. Her research interests include Complex Service Systems Engineering, Service Analytics, and Applied Mathematics.

AHMAD S. ALMADHOR received the B.S.E. degree in computer science from Aljouf University (formerly Aljouf College), Aljouf, Saudi Arabia, in 2005 and the M.E. degree in computer science and engineering from University of South Carolina, Columbia, SC, USA, in 2010 and the Ph.D. degree in electrical and computer engineering at the University of Denver, Denver, CO, USA, in 2019. From 2006 to 2008, he was a Teaching Assistant and College of Sciences manager, then a lecturer from 2011 to 2012, all at Aljouf University, Aljouf, Saudi Arabia. Then, I became a Senior Graduate Assistant and Tutor advisor at the University of Denver in 2013 2019. He is currently an Assistant Professor of CEN and VD of Computer and Information Science College at Jouf University, Saudi Arabia. His research interest includes AI, Blockchain, Networks, Smart and Microgrid cyber security, and integration. Image processing, Video Surveillance systems, PV, EV, Machine, and Deep learning/ Dr. Almadhor's awards and honors include the Aljouf University Scholarship (Royal Embassy of Saudi Arabia in D.C.), Aljouf's Governor Award for excellency, and several others.

ZUNERA JALIL is an Assistant Professor at the Department of Cyber Security, Faculty of Computing & Artificial Intelligence, Air University, Islamabad, and senior researcher at National Cybercrimes and Forensics Lab, National Center for Cyber Security, Islamabad, Pakistan. She earned her Ph.D. in Computer Science with a specialization in Information Security from FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2010. She received her Master's degree in Computer Science in 2007 with a scholarship from the Higher Education Commission of Pakistan. She has served as a full-time faculty member at International Islamic University, Islamabad; Iqra University, Islamabad; and Saudi Electronic University, Riyadh, Saudi Arabia. Her research interests include but are not limited to computer forensics, machine learning, criminal profiling, software watermarking, intelligent systems, and data privacy protection.

ROUBA BORGHOL is an Assistant Professor of Mathematics at Rochester Institute of Technology Dubai. She received a Master's Degree in Applied Mathematics from the University of Claude Bernard II, Lyon, and a Ph.D. in Mathematics from the University of Tours in France in December 2005. Dr. Rouba was employed by the College of Applied Science and Dhofar University as an Assistant Professor for the academic years 2010-2013. The Lebanese University also employed her- Lebanon as an Assistant Professor for the academic year 2008-2009 and as a lecturer at the University of Tours in France during 2005-2007. She was a research fellow at the polytechnic school of Palaiseau in France in 2008. Throughout her fifteen years in academia, she has taught several courses and topics, such as pure mathematics, and applied for mathematics courses for both undergraduate and graduate programs.