IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 5, 2022, accepted April 7, 2022, date of publication April 11, 2022, date of current version April 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3166602

A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics

AHMED ABBASI[1], ABDUL REHMAN JAVED[2], (Member, IEEE), AMANULLAH YASIN[3], ZUNERA JALIL[2], NATALIA KRYVINSKA[4], AND USMAN TARIQ[5]

[1]Faculty of Computing and AI, Air University, Islamabad 44000, Pakistan
[2]Department of Cyber Security, Air University, Islamabad 44000, Pakistan
[3]Department of Creative Technologies, Air University, Islamabad 44000, Pakistan
[4]Information Systems Department, Faculty of Management, Comenius University in Bratislava, 82005 Bratislava, Slovakia
[5]College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia

Corresponding author: Abdul Rehman Javed (abdulrehman.cs@au.edu.pk)

ABSTRACT With the emergence of new digital technologies, a significant surge has been seen in the volume of multimedia data generated from various smart devices. Several challenges for data analysis have emerged to extract useful information from multimedia data. One such challenge is the early and accurate detection of anomalies in multimedia data. This study proposes an efficient technique for anomaly detection and classification of rare events in audio data. In this paper, we develop a vast audio dataset containing seven different rare events (anomalies) with 15 different background environmental settings (e.g., beach, restaurant, and train) to focus on both detection of anomalous audio and classification of rare sound (e.g., events—baby cry, gunshots, broken glasses, footsteps) events for audio forensics. The proposed approach uses the supreme feature extraction technique by extracting mel-frequency cepstral coefficients (MFCCs) features from the audio signals of the newly created dataset and selects the minimum number of best-performing features for optimum performance using principal component analysis (PCA). These features are input to state-of-the-art machine learning algorithms for performance analysis. We also apply machine learning algorithms to the state-of-the-art dataset and realize good results. Experimental results reveal that the proposed approach effectively detects all anomalies and superior performance to existing approaches in all environments and cases.

INDEX TERMS Audio forensics, audio analysis, anomaly detection, key feature extraction, feature selection, machine learning.

I. INTRODUCTION

Technological advancements the world has seen during the past decade, the volume of digital media data on the internet has nearly quadrupled [1], [2]. Smartphones have enabled people to record and store every aspect of their lives in the form of multimedia files [3]–[5]. Moreover, surveillance cameras for monitoring streets, offices, and traffic for security have increased [6]. This exponential increase in multimedia data has called for the need for multiple techniques to analyze this data to be managed and utilized to the best of their ability. Anomaly detection refers to the difficulty of finding patterns in data that do not conform to expected behavior [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

The significance of anomaly detection is because anomalies in data translate to vital information in broad categories of application domains [8]. In the audio category, anomaly detection has several critical applications [9]. For instance, detecting abnormal activities/events in audio can conscientiously supplement video-based methods [2], anomaly detection for machines by analyzing their sounds could be extremely valuable to detect an abnormal performance of the machines in advance [10], to detect abnormal situations that may represent a risk for the public security [11].

With the arrival of smart video surveillance systems, innovative ways for quickly and effectively detecting malicious occurrences or behaviors in monitored settings based on real-time analysis of multimedia streams have emerged [12], [13]. Most real-world audio recordings are complicated in

that they are composed of sequences of many different sounds [9], [14], [15]. If a comparably short sequence of sounds can be distinguished by a human regardless of the acoustic context in which it occurs, it can be classified as an anomaly. For example, the sound of a gunshot in a beach environment or the sound of screams in an office environment. Detecting such anomalies has been researched by researchers during the past few years. Researchers employed a variety of machine learning and deep learning algorithms [16]. However, their signal-to-noise ratios were poor, or their accuracies were low [11], [17].

A. MOTIVATION

Law enforcement and private investigators may learn a lot from a media forensics expert [18]. The audio forensic investigator must be able to detect events from audio in a short time; for example, in a gunshot incident, the gun may not be readily noticeable on the scene or camera, but if there is a gunshot event occurring in audio, so the proposed system should be able to efficiently detect the gunshot event, which helps the forensic investigator during the investigation. In addition, there is a real need for a system that detect abnormal events related to the occurrence of environmental burst like sounds (such as Screams, gunshots, glass break, explosion) that may have been considered ''anomalous'' for the observed environment, thereby diverting the attention of human security operators to a potentially dangerous situation. Furthermore, the authors highlighted the issue of public security, which may be addressed by identifying anomalous noises such as (e.g., footsteps, police siren, baby crying, scream) so that changing the focus of surveillance/human security operator to the specific situation to avoid further harm.

B. CONTRIBUTIONS

This study provides the following contributions to successfully and efficiently identify abnormalities in audios with varied background environments.

- Present a multi-modal open dataset for anomaly detection and rare event classification from audio. For the time being, the dataset consists primarily of rare events with 15 background audios. The final collection of the dataset comprises seven different types of rare events—baby cries, gunshots, broken glasses, footsteps, police sirens, explosions, and screams—that have been artificially blended with background audio recordings from 15 various environmental contexts (i.e., office, Library, and Park). Detailed descriptions of the dataset are presented in the following Section III.
- Utilize a supreme feature extraction technique to extract MFCC features from the audio signals and Principal Component Analysis (PCA) for the selection of suitable features for anomaly detection in the audio signal.
- Propose a practical machine learning approach for anomaly detection and classification of rare events in audio data embedded in various types of background sounds.
- Analyze and validate the effectiveness of those feature extraction and feature selection approaches on the performance of the machine learning algorithms for anomaly detection and present a comparative analysis of the suggested approach with other state-of-the-art studies, which effectively enhances the detection rate with consistent performance.

C. ORGANIZATION

The structure of this paper is as follows. The section II addresses related work. The dataset used for testing and early analysis is discussed in section III. The proposed technique for anomaly identification and categorization of unusual occurrences in audio data is described in Section IV. Section V articulates the experimental setup and findings. Section VI contains a discussion, and Section VII provides the conclusion and future work.

II. LITERATURE REVIEW

Several methods, primarily based on AI/ML-based methodologies, have been used to detect abnormalities throughout the last decade. In [19], the authors employed a technique that reconstructs the features for anomaly identification based on an LSTM-based network that detects abnormalities from subsampled signals. The authors of [20] used non-uniform sampling for audio subsampling to make low-volume samples that include higher frequencies than Nyquist. The LSTM-based auto-encoder network is then used for anomaly detection, in which the signal is made demultiplex and accepted as input from the endpoint. The authors in [21] used Convolutional Autoencoder (CAE). The CAE is used to detect abnormal activities from the audio that are overlaid with natural factory soundscapes. The CAE-based approach gives better results than One-Class Support Vector Machines. They used a limited number of audios in their experimental work. The authors in [22] used sequence-to-sequence autoencoder models on audio features extracted from the streaming audio signals. They found that Convolutional Long Short-Term Memory autoencoders perform better than sequential Convolutional autoencoders under diverse signal-to-noise ratio conditions of audio events.

In [23] this paper, the authors used two models to achieve the goal. In [24], the authors used a modular deep convolutional autoencoder with a dense bottleneck structure for unsupervised anomaly detection. They also applied Maximum Mean Discrepancy (MMD). To efficiently learn the features, they used MMD. For training, the authors employed two models. The first is a 1D-convolutional-encoder, and the second is the WaveNet-decoder model. The identical encoder/decoder structures are trained to learn a mapping function between different mel-scaled frequency bands. An SVM model is trained to predict anomalies and examine the latent space representation learned by the autoencoders. They found that this method paves the way toward

semi-supervised or self-supervised training for detecting anomalies.

The authors of [25] suggested a methodology for using Huffman coding. This approach is utilized for anomaly detection in audio to obtain benefits such as variable event length and reduced reliance on cluster information, and it was discovered that this method enhances outcomes with little computing overhead. In [26] the authors introduced a training strategy, primarily used in unsupervised ADS. The authors suggested a batch uniformization technique. First, they reduced the weighted mean score. Here weight is defined as the reciprocal of each sample's probability density. The authors found that this method is appropriate for an unsupervised anomaly detection system based on a deep neural network (DNN).

The authors of [27] suggested an auto-encoder that leverages the residual error, which represents reconstruction quality, to find the anomaly. In [28], the authors employed an auto-encoder model to detect the anomalies in audios. The audios were recorded in home surroundings. The main limitation of this work is that they used a very less number of audio events and background audios for training and testing. In [17], the authors adopt WaveNet architecture model. This model was developed for raw audio synthesis, ADA, and significant performance increases over deep-convolutional-autoencoders (DCA). The WaveNet model outperformed the DCA technique; however, it earned a relatively low AUC ROC score overall, indicating that the model did not perform well on the dataset. Table 1 tab summarises the literature review.

III. NETWORK MODEL, DATASET AND PRELIMINARIES

We evaluate the suggested system's performance for an automated surveillance application that should be capable of identifying the following occurrences (called ''abnormal'' or ''anomaly'' in the observed environment): baby cries, gunshots, broken glasses, footsteps, police siren, explosions, and screams. We create the dataset by mixing different rare events with 15 background audio datasets fetched from the TUT Acoustic Scenes 2016 dataset [29]. A Soundman OKM II Classic/Studio A3 head-microphone and an R-09 from Edirol/Roland wave recorder were used for the TUT Acoustic Scenes 2016 audio dataset recordings. The recording quality is excellent. The TUT Acoustic Scenes 2016 collection is made up of real-world audio recordings. The recorded audio is remarkably comparable to the sound that reaches the human wearing the equipment's human auditory system. The final collection of the dataset consists of seven types of unusual events—baby cries, gunshots, broken glasses, footsteps, police sirens, explosions, and screams. These audio events were then synthetically mixed with the 15 background environmental contexts audios(beach, bus, home). The final created dataset is available at https://www.kaggle.com/ahmedabbasi/audioanomalydataset. The datasets utilized in this experiment are summarised in Figure 1. Existing techniques concentrated on identifying only one type of audio (only a rare event or a background sound), resulting in poor performance throughout the test under actual situations. As a result, the aim of extending the dataset is:

- To focus on detecting anomalous audio and classifying rare sound events.
- To focus on using the audio information for surveillance purposes.
- To encourage other researchers in the field to use this dataset for testing their methods for anomaly detection and rare event classification.

Rare events are randomly mixed at different ''event-to-background'' ratios. The original audio mixtures are sampled at 44.1KHz with a 24-bit resolution. The data set contains highly noisy environmental sounds, making event detection more difficult in some environments and challenging the detection and classification of events.

$$B_j(n) = \sum_{i=1}^{n} bg_i(n) \tag{1}$$

We begin by gathering 1170 background audio recordings from the TUT 2016 dataset. First, the background sound $bg_i(n)$ is selected randomly by defined number of audios n $\in \{1, 2, 3, \ldots, \}$ as mentioned in equation 1. $bg_i(n)$ are the ''n'' carefully chosen background audios that are utilized to produce the complex environmental sound by combining several unusual events.

$$y_j(n) = \sum_{e=1}^{N_e} \oplus_{[N_e, B_j]} B_j(n) \tag{2}$$

Once the background audio has been selected, a number N_e of rare events is randomly chosen and superimposed on the background audio. As a result, the unusual occurrence might be present in the final data set and appear with different background audio each time.

In equation 2, with $\oplus_{[N_e, B_j]}$ we define an operator that mixes the rare events N_e with the background audio $B_j(n)$ at random positions of audio signal. The final dataset consists of 8,922 audios, and the total duration is about 75h making the database huge. We split the final mixture of audio files of 30 seconds into two partitions. We employed the first split for training and the second portion for subsequent assessment. Figure 2 depicts sound waves of anomalous sounds for all types of rare events.

IV. PROPOSED APPROACH

The proposed approach comprises multiple steps data analysis, feature extraction, feature reduction, processing data, and finally, detection of anomalous audio and classification of rare sound events as shown in Figure 3. The data analysis involved the visualization of audio waveform and spectrogram to extract meaningful insights from audio. The feature extraction techniques use a featured ensemble of MFCC,spectral_rolloff features, spectral_centroid features, spectral_contrast features, spectral_bandwidth features, and

TABLE 1. Literature review overview.

Reference	Approach	Limitation
[28]	Adversarial autoencoders	Limited anomaly events and backgrounds
[17]	Adapt WaveNet	Lower Accuracy
[21]	Convolutional Autoencoder (CAE)	Limited dataset and number of audios
[22]	Convolutional Long Short-Term Memory autoencoders	Lower Accuracy

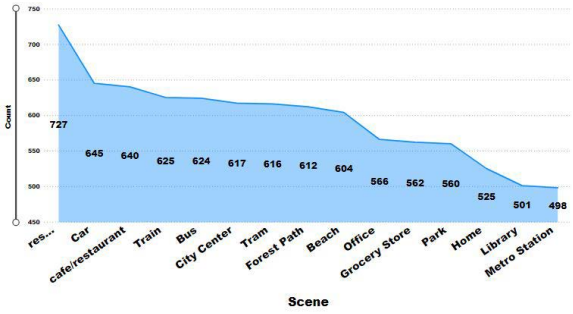

FIGURE 1. Dataset overview.

zero_crossing_rate features. We used the feature reduction technique because feature ensemble has useful and non-useful features, so we need to remove these non-useful features and pass only useful features to machine learning classifiers for better classification results. We used Principal Component Analysis (PCA) for feature reduction, which selects the most appropriate features that are finally passed to the classifiers for classification and anomaly event detection. An audio surveillance system must recognize unusual occurrences even when blended with a variety of background audios of varying energy levels—as a result, training a model with a collection of the training set, which contains only one sort of audio (at a time, either an exceptional event or a background sound) would lead to a performance in the test phase in pragmatic scenarios. We decided to create a train and test set where the different audios are already layered rather than separated to address this issue. Additionally, the proposed approach can efficiently detect anomalous events in audio, which helps the forensic investigator for further investigation.

We have performed supreme feature engineering techniques, which results in the models' ability to detect anomalies exceptional. After gathering many background audios from the TUT challenge 2016, we combined them with events of interest in various ways to get a large data set. To provide very challenging event detection tasks, the data set comprises loud environmental noises, such that events may be more challenging to detect in specific situations, attempting to make event detection and categorization extremely difficult. The audio clips are divided into two distinct divisions, each containing 80% and 20% of the total sounds from the original collection. The audios from the first partition were used to create the training set, while the audios from the second portion were utilized to create the test set.

A. PRE-PROCESSING

Pre-processing is necessary to acquire remarkable performance in any machine learning model before training the

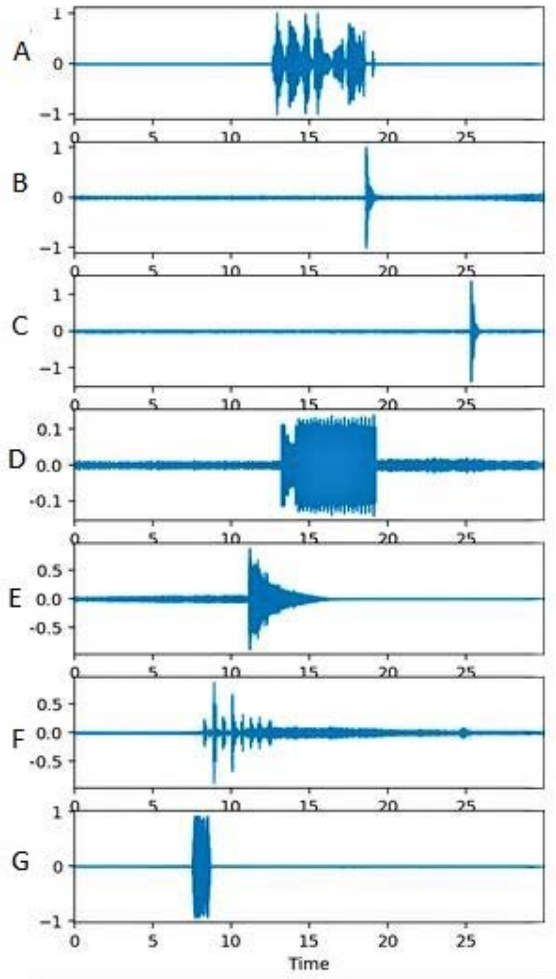

FIGURE 2. Waveform of anomalous audios. Key: Baby Cry-A, Gunshot-B, Glass Break-C, Police-D, Explosion-E, Foot Steps-F, scream-G.

classifiers. Therefore, audio data includes a handful of pre-processing procedures that must be taken before it is delivered for further analysis. The first stage is data framing, which involves converting the audio data into a machine-readable format. We acquire values after a particular time. For example, in a 10-second audio file, we extract values every second, which is audio data sampling, and the sampling rate is the rate at which it is sampled. In our case, by default, the sampling(frame) rate is 44100. This sampling(frame) rate is the frame values of an audio file within 1-s and calculates the overall frames by multiplying the sampling(Frame) rate by the time of an audio file as explicated in Eq. 3.

$$totalframe = samplingrate \times time \tag{3}$$

In our case, if an audio file "file1" has a 30-s time, then the total frame rate of this file can be calculated

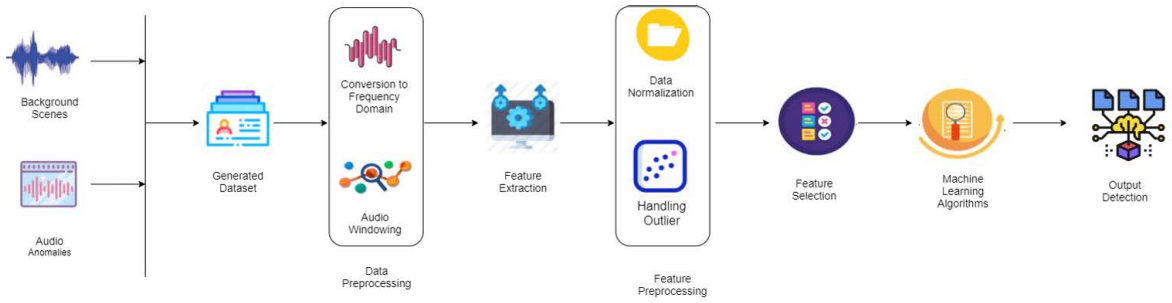

FIGURE 3. Visual representation of proposed approach for detection of anomalous audio and classification of rare sound events.

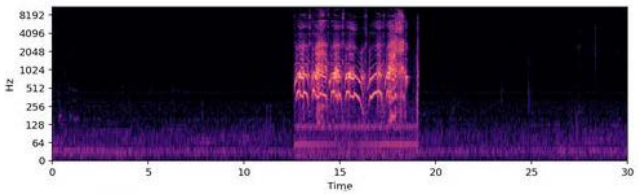

FIGURE 4. Spectrogram of audio signal.

by the formula in Eq. 4.

$$file1 = 44100 \times 30 = 1323000 \tag{4}$$

Data framing is used to fix the sampling(frame) rate of each audio file [30]. The audio processing procedure begins with extracting key acoustical characteristics, preceded by decision-making techniques involving detection, classification, and knowledge fusion. In the following phase, we represent audio data by transforming it into a new data representation domain, the frequency domain. We needed a lot more data points to represent the entire audio data when we sampled it, and the sampling rate should be as high as feasible. Each sample represents the amplitude of the audio waveform at a certain time interval. To visually inspect the audio signal, we create a spectrogram. It shows the signal intensity, or "loudness," across time at various frequencies contained in a certain waveform, as seen in Figure 4. It shows a spectrogram of a baby cry audio waveform. The vertical axis displays frequencies ranging from 0 to 8kHz, while the horizontal axis displays the duration of the audio clip. In a spectrogram, purple colors represent the amplitude of a sound wave.

We choose a standard scaler for feature normalization. Standard scaling is utilized in this study to standardize data within certain ranges (e.g., 0 and 1). The goal of Standard Scaler is to rescale features so that they are roughly standard normally distributed. We utilize a conventional scaler to modify the data such that it eliminates the mean and scales each feature/variable to unit variance, as shown in Eq. 5, where y is our standardized form of x.

$$y = \frac{(x - \mu)}{\sigma} \tag{5}$$

B. FEATURE EXTRACTION

Every audio signal contains a variety of characteristics/features. We must, however, extract the features related to the event that we will detect. We employed Mel-frequency Cepstral Coefficients for this purpose (MFCC).

TABLE 2. Features extracted.

Feature Group	Features in group
Cepstral	MFCC 0 – 39
Spectral	Roll-off point, Centroid, Contrast, Bandwidth
Raw Signal	Zero Crossing Rate
Signal Energy	Root Mean Square

The MFCC is a feature extraction method, and in this study, we used 39 MFCC features. In sound processing, MFCCs features are the most often utilized in speech recognition [31]. In this work, MFCC is exploited for anomaly detection. After the pre-processing of anomaly audio signals, the MFCC vector will be extracted from each frame of the audio waveform in the form of a vector group. This study uses MFCCs, spectral_rolloff, spectral_centroid, spectral_contrast, spectral_bandwidth and zero_crossing_rate features for experimentation. We construct these features by taking the mean and standard deviation of values computed at each frame and combining them to get the value for the relevant feature. FIGURE 5 depicts the MFCC series of infant cry audio files by converting the audio waveform into the frequency domain using the Fourier transform of a signal, then mapping the powers of the spectrum produced onto the mel scale. After that, compute the discrete cosine transform of the list of mel log powers by taking the logs of the powers at each of the mel frequencies. The MFCCs are the resultant spectrum's amplitudes. The features collected from each feature group are described in Table 2. Each audio file has 270 characteristics extracted, and the results are saved in a data frame. We selected only suitable features and removed all non-useful features using Principal Component Analysis [32]. In the end, 65 most important features passed to models for anomaly detection. To evaluate the usefulness of selected features, we calculate the explained_variance_ratio of PCA. The 97% value of explained_variance_ratio shows that the selected data is valuable.

C. CLASSIFICATION MODELS, PARAMETER SETTING, AND ALGORITHM

We employ the Support Vector Machine (SVM), K-Nearest Neighbor Algorithm (KNN), Extreme Gradient Boosting (XGB), Multi-layer Perceptron (MLP), Random Forest (RF),

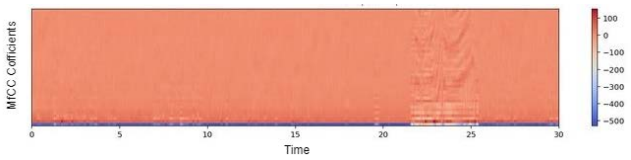

FIGURE 5. Visualize the MFCC features.

and Logistic Regression (LR) machine learning techniques to detect anomalies and classify unusual occurrences, as well as to assess the efficacy of our suggested approach.

Algorithm 1 shows the working of the proposed approach. Let D indicate the dataset that contains the instance $I = \{i_1, i_2, \ldots, i_n\}$ and t is the time duration of audio. F represents the data framing to get meaningful information from the time series audio signal by multiplying the sampling rate (SR) with $time(t)$. For a visual representation of the audio, the first step is to convert the time series audio signal M to frequency domain $x(n)$. $X(m, w)$ is essentially the Fourier transform of $x[n]w[n-m]$, a complicated function describing the signal's phase and amplitude with time and frequency, Where $w(n)$ is the window function, commonly a Hann window, and $x(n)$ is the signal to be transformed (note the difference between the window function w and the frequency w). $X(m, w)$ here m is time, discrete, but w is the frequency and continuous. In the last step, STFT $\{x[n]\}(m, w)$ is the visual representation of the signal strength, or ''loudness,'' of a signal over time at various frequencies. In our case, we have frequencies from 0 to 8KHz. After the power spectrogram, the filter bank processing is carried out on the power spectrum using mel-scale S_k to extract the valuable features. Eventually, the 270 MFCCs are calculated, where k is the number of mel cepstrum coefficients, S_k is the output of filterbank, and C_n is the final MFCC coefficients. The feature vector is formed in the next step by calculating the mean and standard deviation of values determined by each frame and storing them in the data frame (df). For selecting convenient features, the first step is to standardize the data.

A standard scaler is used to convert the data into numeric, where the mean is removed, and each feature is scaled to unit variance. The second step is to calculate the covariance matrix of the df. The dimension of the covariance matrix is represented as n ∗ p. The third step is to calculate eigenvalues and eigenvectors. The dimension of eigenvectors is represented as p ∗ m. Finally, X_PCA represents 65 suitable features for experimentation. The features are chosen, and the resulting feature vectors are put into machine learning models. The model learns the anomaly sequence patterns. The properly chosen characteristics increase the learning operations and aid in achieving greater accuracy in the anomaly detection process. The last phase is model prediction, used to discover abnormalities in new data.

V. EXPERIMENTAL ANALYSIS AND RESULTS

This study proposes a generic system by extending the dataset for anomaly detection and classification of a rare event in real-life audio. We conduct an experimental evaluation of the

Algorithm 1 Anomaly Detection and Classification

Input: data ← Audio Data
Output: Anomalies Detection and Classification
1: $F \leftarrow SR * t(audio)$ {Data Framing}
2: $M \leftarrow F(audio)$ {Convert audio signal from time to frequency domain}
3: $x(n) \leftarrow M$ {Windowing audio signal}
4: X(m,w) ← $\sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn}$ {FFT}
5: $STFT\ \{x[n]\}(m, w) \leftarrow X(m, w)$ {Display Spectrogram}
6: $S_k \leftarrow 2595x \log_{10}(1 + f/700)$ {Calculate mels for any frequency}
7: $C_n \leftarrow \sum_{n=1}^{k} (\log S_k) \cos[n(k - 1/2)\pi/k]$ {MFCC features}
8: get_mfcc ← $\frac{1}{n} * np.sum(np.square(C_n))$ {Mean of MFCCs}
9: $get\_mfcc \leftarrow \sqrt{\frac{\sum_{i=1}^{n}(C_n-\mu)^2}{N}}$ {st.deviation of MFCCs}
10: $df \leftarrow get\_mfcc$ {store features in new df}
11: $df \leftarrow \frac{(df-\mu)}{\sigma}$ {Apply standard sclaer}
12: $Y \leftarrow DataMatrix(n * p)(df)$ {calculate Covariance Matrix of df}
13: $V \leftarrow EigenVectorMatrix(p * m)(df)$ {Calculate eigen values and eigen vectors.}
14: X_PCA ← Y.V {Top k featrues selected}
15: $acc \leftarrow []$
16: $MLModels \leftarrow [SVM, KNN, XGB, MLP, RF, LR]$ {Model Training}
17: **for** each i in MLModels **do**
18: $acc \leftarrow getClassification(i)$
19: $i \leftarrow i++$
20: Evaluate Accuracy
21: Evaluate Precision, Recall, F-Score, Roc Curve and Confusion Matrix
22: return output
23: **end for**{Anomaly Detection}
24: TL ← Predict Class(X_PCA)
25: **for** i in range(1, len(TL)) **do**
26: **if** (TL[i] = y_test[i]) **then**
27: return TL[i]
28: **else**
29: return y_test[i]
30: **end if**
31: **end for**

suggested technique using a typical audio surveillance application in which seven types of audio events must be detected: baby cry, gunshot, glass breaking, footsteps, explosion, and scream. The experiments of this study were carried out utilizing different machine learning techniques. Six machine learning models are used for experiments, including Random Forest, KNN, XGB, MLP, SVM, and logistic regression. We used Google Colab for experimental implementation. Following the experimentation, the findings are compared to the other existing state-of-the-art approaches depicted in Figure 6. Accuracy, precision, recall, and F-score are the performance evaluation metrics. We conduct experiments on google colab with the windows 10 professional operating system. The CPU is Intel Xeon Processor. Furthermore, GPU is Tesla K80. To conduct the experimental assessment, the dataset is divided into two disjoint groups: training and testing, which account for 80% and 20% of the total number of audios in the newly created dataset, respectively. 80% of the data is used for training, while 20% is used for testing.

TABLE 3. AUC scores of baseline models on each dataset.

Scene	Baseline CAE %	Baseline WaveNet %
Beach	69	72
Bus	79	83
cafe/restaurant	69	76
Car	79	82
City Center	75	82
Forest Path	65	72
Grocery Store	71	77
Home	69	69
Library	59	67
Metro Station	74	79
Office	78	78
Park	70	80
residential area	73	78
Train	82	84
Tram	80	87

A. BASELINE CAE AND WaveNet MODEL

We employ a convolutional autoencoder (CAE) and the WaveNet model as a baseline. We compare the results of our machine learning model with these two baseline model results. The CAE model has 20 layers. Ten are encoder layers, and the remaining 10 are decoder layers. The WaveNet model comprises 20 layers, but it comprises two stacks of 10 convolutions. They used the dataset named DCASE challenge Task 2 published in 2017 [33]. This dataset includes three unusual events (baby cry, glass break, and gunshot) and background audio from 15 different environmental situations (e.g., Bus, Forest Path, and Home). CAE and WaveNet models are trained on training sets and evaluated on test sets. During the testing phase, they calculated the Area Under the ROC Curve (AUC) to demonstrate the models' capacity to identify abnormalities. TABLE 3 shows the performance of both models throughout the 15 datasets, with a tie in the home and office settings. The proposed approach achieves better results than the baseline.

B. RESULTS

The proposed system's performance is assessed using a huge dataset of audio samples that includes seven unique, unusual occurrences (anomalies) that are artificially blended with fifteen diverse background audios. ADS, the suggested technique, is utilized for anomaly identification and categorization of unusual occurrences in real-world audio. Five evaluation metrics were used in this proposed approach. First is accuracy, then precision, recall, F1-score, and at the end, the ROC curve was a plot to evaluate the model ability on the given dataset.

Tables 4 and 5 show the experimental outcomes of the machine learning models. Table 4 displays the machine learning models' accuracy, precision, recall, and F1-score, whereas Table 5 displays the ROC curve score of the machine learning models across the 15 datasets. According to Table 4, MLP model performed quite well on average. On the café environment dataset, the MLP model obtained the highest accuracy score of 99.08%. Furthermore, the MLP model got the maximum precision, recall, and F1-score on the cafe environment dataset, 99.04, 99.04, and 99.03, respectively.

TABLE 4. Classifier performance (%) to detect anomalies. Key: Residential Area- RA, City Center-CC, Forest Path-FP, Grocery Store-GS, Library-Lib, Metro Station-MS, Office-Off.

Scene	Results	SVM	KNN	XGB	MLP	RF	LR
Beach	Accuracy	95.05	90.01	93.02	94.01	91.01	95.01
	Precision	96.01	91.02	92.01	94.03	92.01	95.02
	Recall	95.01	90.03	92.01	93.03	91.02	94.01
	F1-score	95.01	90.02	92.02	93.03	91.02	94.02
Bus	Accuracy	98.11	87.73	97.15	99.05	95.28	96.22
	Precision	98.03	89.03	97.01	99.02	97.03	95.01
	Recall	98.03	88.02	97.01	99.03	96.01	94.03
	F1-score	98.01	88.02	97.03	99.01	96.02	94.02
Cafe	Accuracy	95.41	91.74	97.24	99.08	97.24	99.01
	Precision	95.01	92.02	96.03	99.04	96.03	99.01
	Recall	93.02	92.02	96.01	99.04	95.03	99.01
	F1-score	93.01	92.01	96.02	99.03	95.02	99.01
Car	Accuracy	98.13	90.65	97.19	98.13	99.06	98.13
	Precision	98.02	91.03	96.03	98.02	99.01	98.01
	Recall	98.01	91.12	96.02	98.01	99.01	98.02
	F1-score	98.01	91.13	96.01	98.02	99.01	98.01
CC	Accuracy	94.17	89.32	92.23	97.08	94.17	95.14
	Precision	94.12	92.05	92.12	96.13	94.03	95.03
	Recall	94.12	89.07	91.12	96.13	94.02	95.03
	F1-score	94.12	90.06	91.11	96.12	94.02	95.02
FP	Accuracy	99.03	98.07	98.02	98.02	98.03	95.19
	Precision	99.12	98.10	98.03	98.02	98.02	96.01
	Recall	99.12	98.13	98.02	98.02	98.02	95.01
	F1-score	99.12	98.11	98.02	98.02	98.03	95.01
GS	Accuracy	93.47	93.47	91.30	97.82	93.47	94.56
	Precision	94.03	94.03	91.01	98.03	94.01	96.02
	Recall	93.02	93.01	91.01	98.02	93.03	95.01
	F1-score	93.01	93.02	91.02	98.01	93.01	94.02
Home	Accuracy	91.95	86.20	87.35	97.70	90.80	94.25
	Precision	93.12	87.01	88.03	97.10	89.03	95.12
	Recall	92.11	86.02	87.03	97.11	89.02	94.11
	F1-score	92.11	86.02	88.02	97.10	88.01	94.11
Lib	Accuracy	96.25	95.01	95.01	98.02	97.50	98.75
	Precision	97.01	97.02	95.03	99.01	98.02	99.03
	Recall	96.01	95.02	95.03	99.01	97.02	99.03
	F1-score	96.03	95.02	95.01	99.02	98.01	99.03
MS	Accuracy	96.25	81.25	95.02	96.25	90.02	97.50
	Precision	96.12	85.13	96.14	97.13	91.12	97.13
	Recall	96.12	81.13	95.14	96.13	89.12	97.13
	F1-score	96.11	82.12	95.12	96.12	89.11	97.12
Off	Accuracy	97.80	98.90	97.80	98.03	98.03	96.70
	Precision	98.12	99.12	97.10	97.03	98.03	97.03
	Recall	98.11	99.11	97.10	99.02	99.02	97.03
	F1-score	98.11	99.11	97.09	97.02	98.02	97.03
Park	Accuracy	96.66	90.01	95.55	98.80	95.55	94.44
	Precision	97.03	92.01	96.13	99.04	95.14	94.13
	Recall	97.03	90.01	96.12	99.03	94.12	94.12
	F1-score	97.02	90.02	96.10	99.03	94.12	94.12
RA	Accuracy	97.01	90.04	92.02	96.08	95.19	97.61
	Precision	97.01	91.01	93.02	96.04	96.03	98.10
	Recall	97.01	90.01	91.02	96.04	95.03	98.10
	F1-score	97.01	91.01	91.02	96.03	95.02	98.11
Train	Accuracy	95.01	80.03	90.47	95.01	91.42	95.23
	Precision	95.01	82.02	91.03	95.01	91.02	96.03
	Recall	95.03	80.02	90.01	95.03	90.02	95.01
	F1-score	950.2	80.01	90.03	95.02	90.01	95.03
Tram	Accuracy	98.01	90.09	97.02	99.01	95.04	94.05
	Precision	98.02	91.04	97.05	99.02	94.04	94.05
	Recall	98.02	90.05	97.04	99.05	94.02	94.04
	F1-score	98.05	90.02	97.04	99.02	94.04	94.05

Furthermore, Table 5 displays the AUC values for all anomaly detection datasets. MLP model outperformed all other machine learning models on the cafe environment dataset in terms of AUC.

C. COMPARATIVE ANALYSIS WITH BASELINE APPROACHES

This study compares the findings of the suggested technique to another state-of-the-art study [17], whose experimental

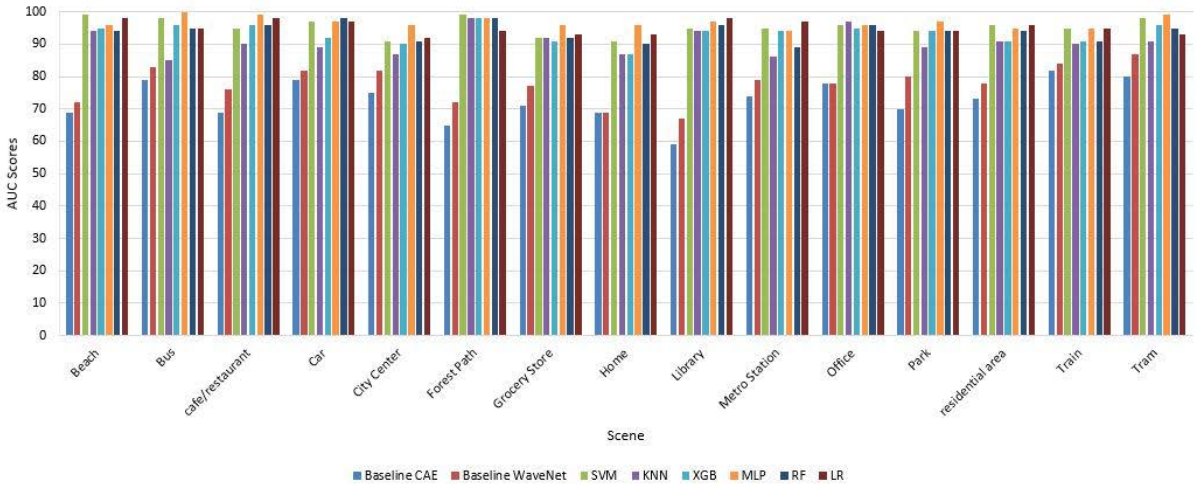

FIGURE 6. Comparison of the proposed approach with baseline CAE, WaveNet model and conventional machine learning algorithms.

TABLE 5. AUC scores on all anomaly detection datasets.

Scene	SVM	KNN	XGb	MLP	RF	LR
Beach	0.99	0.94	0.95	0.96	0.94	0.98
Bus	0.98	0.85	0.96	1.00	0.95	0.95
cafe/restaurant	0.95	0.90	0.96	1.000	0.96	0.98
Car	0.97	0.89	0.92	0.97	0.98	0.97
City Center	0.91	0.87	0.90	0.96	0.91	0.92
Forest Path	0.99	0.98	0.98	0.98	0.98	0.94
Grocery Store	0.92	0.92	0.91	0.96	0.92	0.93
Home	0.91	0.87	0.87	0.96	0.90	0.93
Library	0.95	0.94	0.94	0.97	0.96	0.98
Metro Station	0.95	0.86	0.94	0.94	0.89	0.97
Office	0.96	0.97	0.95	0.96	0.96	0.94
Park	0.94	0.89	0.94	0.97	0.94	0.94
residential area	0.96	0.91	0.91	0.95	0.94	0.97
Train	0.95	0.90	0.91	0.95	0.91	0.96
Tram	0.98	0.91	0.96	0.99	0.95	0.93

TABLE 6. AUC score gain comparison of *Proposed* with baseline approach.

Scene	Highest Baseline (%)	Proposed (%)	Gain (%)
Beach	72	99	27
Bus	83	100	17
Cafe/restaurant	76	99	23
Car	82	98	16
City Center	82	96	14
Forest Path	72	99	27
Grocery Store	77	96	19
Home	69	96	27
Library	67	98	31
Metro Station	79	97	18
Office	78	97	19
Park	80	97	17
Residential area	78	97	19
Train	84	96	12
Tram	87	99	12

circumstances are similar to the settings used in this study. The performance of both the baseline approach and proposed approach across the 15 datasets is shown in FIGURE 6. The ROC curves, which provide an overall evaluation of classification performance, are used to evaluate the performance of both approaches (baseline and proposed). The suggested method consistently outperforms the baseline CAE and WaveNet models in virtually all datasets, with corresponding scores closer to 1. For a perfect classification, we consider the ROC curve score to be closer to 1. The greater the value of this metric, the better the overall performance of the suggested system. An audio surveillance system must identify events even when blended with various background audios of varying energy levels. As a result, we suggested a general approach for anomaly identification and categorization of a rare event in real-life audio by expanding the dataset and employing machine learning. The baseline technique comprises training a system by utilizing a set of training samples to identify only (whether an anomaly exists or not), but we concentrated on detecting anomalous audio and categorizing rare sound events. Because the original dataset comprises very loud ambient noises, events may be more challenging to identify in various situations, making event detection and categorization difficult. In the baseline approach, it is noticeable that the performance of the CAE and WaveNet model differs dramatically across diverse acoustic settings, demonstrating that varied acoustic environments may significantly alter the models' capacity to detect anomalies. TABLE 6 compares our suggested technique to the best performing classifier in the referenced paper. Across all datasets, we find that the suggested strategy outperforms the baseline approach by a substantial margin. The proposed approach gained the highest AUC gain, a score of 27 on the beach, forest path, and home environment dataset. The proposed approach also gained better results for the rest of the classes.

VI. DISCUSSION

This study presents anomaly detection and classification using machine learning algorithms. Terrorism is posing serious and rising threats all across the world. Anomaly detection in audio is critical for detecting events connected to environmental bursts and explosion-like sound occurrences (e.g., gunshots screaming, explosion) that may be regarded as ''odd'' for the observed environment to make our living environment safer. For this purpose, a dataset is created using seven different anomalous sounds and 15 different

background sounds. All the anomalous sounds are then embedded in each different background to create a dataset for experimentation. Then, many features from the dataset are retrieved, and a feature selection technique, Principal Component Analysis, is used to limit the number of features and choose just useful features. For anomaly detection, several machine learning methods are used in the dataset. After thorough experimentation, we found that the MLP machine learning classifier consistently performed well for every anomaly event in each background scenario. Using PCA for feature selection and MLP classifier for detection gives remarkable accuracy that can be very useful when applied in real-world scenarios. As a result, our experimental results demonstrated that the suggested technique detects and classifies abnormalities more effectively than existing state-of-the-art studies.

VII. CONCLUSION

As the velocity of multimedia generation increases, there is a need for techniques to analyze that data. Anomalies mean a deviation from normal or expected behavior. Detection of anomalies can serve as an essential tool for enhancing persons' security and maintaining public and private assets. A customized dataset has been created by mixing rare events with 15 background audios fetched from the TUT Acoustic Scenes 2016 dataset to detect anomalous audio and classify rare sound events. To detect anomalies in audio data, this study conducted experiments using multiple features of audio data. For feature engineering, this study extracts various features from the audio signal and then applies the PCA feature selection technique to select the minimum number of best-performing features for optimum performance. Several machine learning algorithms are employed on the selected feature set to detect seven different anomalous events in 15 different background environments. Experiments demonstrated that our technique outperformed existing state-of-the-art research for anomaly identification in audio data in all circumstances. In the future, we plan to expand our dataset to include a wider variety of anomalies and background scenes and analyze the effectiveness of multiple machine learning algorithms on different types of anomalies.

REFERENCES

[1] A. Abbasi, A. R. Javed, C. Chakraborty, J. Nebhen, W. Zehra, and Z. Jalil, "ElStream: An ensemble learning approach for concept drift detection in dynamic social big data stream learning," *IEEE Access*, vol. 9, pp. 66408–66419, 2021.

[2] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection," in *Proc. DCASE Workshop*, 2020, pp. 51–55.

[3] A. R. Javed, M. O. Beg, M. Asim, T. Baker, and A. H. Al-Bayatti, "AlphaLogger: Detecting motion-based side-channel attack using smartphone keystrokes," *J. Ambient Intell. Hum. Comput.*, pp. 1–14, Feb. 2020.

[4] M. U. Khan, A. R. Javed, M. Ihsan, and U. Tariq, "A novel category detection of social media reviews in the restaurant industry," *Multimedia Syst.*, pp. 1–14, Oct. 2020.

[5] A. R. Javed, S. U. Rehman, M. U. Khan, M. Alazab, and H. U. Khan, "Betalogger: Smartphone sensor-based side-channel attack detection and text inference using language modeling and dense multilayer neural network," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–17, Sep. 2021.

[6] W. Ahmed, A. Rasool, A. R. Javed, N. Kumar, T. R. Gadekallu, Z. Jalil, and N. Kryvinska, "Security in next generation mobile payment systems: A comprehensive survey," *IEEE Access*, vol. 9, pp. 115932–115950, 2021.

[7] A. R. Javed, M. Usman, S. U. Rehman, M. U. Khan, and M. S. Haghighi, "Anomaly detection in automated vehicles using multistage attention-based convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4291–4300, Jul. 2021.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[9] O. K. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Trans. Multimedia*, vol. 23, pp. 3978–3985, 2020.

[10] S. M. Jaile *et al.*, "Anomaly detection using audio signals," M.S. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2020.

[11] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An ensemble of rejecting classifiers for anomaly detection of audio events," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 76–81.

[12] 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," *Eng. Appl. Artif. Intell.*, vol. 106, Nov. 2021, Art. no. 104456.

[13] L. Jing, B. Liu, J. Choi, A. Janin, J. Bernd, M. W. Mahoney, and G. Friedland, "DCAR: A discriminative and compact audio representation for audio processing," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2637–2650, Dec. 2017.

[14] L. Lu and A. Hanjalic, "Audio keywords discovery for text-like audio content analysis and retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 74–85, Jan. 2008.

[15] K. Umapathy, S. Krishnan, and S. Jimaa, "Multigroup classification of audio signals using time-frequency parameters," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 308–315, Apr. 2005.

[16] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 376–380.

[17] E. Rushe and B. M. Namee, "Anomaly detection in raw audio using deep autoregressive networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3597–3601.

[18] A. R. Javed and Z. Jalil, "Byte-level object identification for forensic investigation of digital images," in *Proc. Int. Conf. Cyber Warfare Secur. (ICCWS)*, Oct. 2020, pp. 1–4.

[19] Y. Kawaguchi, "Anomaly detection based on feature reconstruction from subsampled audio signals," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2524–2528.

[20] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[21] T. B. Duman, B. Bayram, and G. İnce, "Acoustic anomaly detection using convolutional autoencoders in industrial processes," in *Proc. Int. Workshop Soft Comput. Models Ind. Environ. Appl.* Cham, Switzerland: Springer, 2019, pp. 432–442.

[22] B. Bayram, T. B. Duman, and G. Ince, "Real time detection of acoustic anomalies in industrial processes using sequential autoencoders," *Expert Syst.*, vol. 38, no. 1, Jan. 2021, Art. no. e12564.

[23] R.-G. Colt, C.-H. Várady, R. Volpi, and L. Malagò, "Automatic feature extraction for heartbeat anomaly detection," 2021, *arXiv:2102.12289*.

[24] I. Thoidis, M. Giouvanakis, and G. Papanikolaou, "Audio-based detection of malfunctioning machines using deep convolutional autoencoders," in *Proc. Audio Eng. Soc. Conv.*, vol. 148. New York, NY, USA: Audio Engineering Society, 2020.

[25] P. Kumari and M. Saini, "Anomaly detection in audio with concept drift using adaptive Huffman coding," 2021, *arXiv:2102.10515*.

[26] Y. Koizumi, S. Saito, M. Yamaguchi, S. Murata, and N. Harada, "Batch uniformization for minimizing maximum anomaly score of DNN-based anomaly detection in sounds," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 6–10.

[27] D. Y. Oh and I. D. Yun, "Residual error based anomaly detection using auto-encoder in SMD machine sound," *Sensors*, vol. 18, no. 5, p. 1308, Apr. 2018.

[28] E. Principi, F. Vesperini, S. Squartini, and F. Piazza, "Acoustic novelty detection with adversarial autoencoders," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3324–3330.

[29] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1128–1132.

[30] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Heartbeat sound signal classification using deep learning," *Sensors*, vol. 19, no. 21, p. 4819, Nov. 2019.
[31] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Syst., Signal Process.*, vol. 40, no. 1, pp. 466–489, Jan. 2021.
[32] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Mar. 2018, pp. 379–383.
[33] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Workshop Detection Classification Acoust. Scenes Events (DCASE)*, 2017, pp. 1–9.

AHMED ABBASI is currently pursuing the master's degree in data science with Air University, Islamabad, Pakistan. He is currently with the Department of Cyber Security, Air University, where he is also working with the National Cybercrimes and Forensics Laboratory.

ABDUL REHMAN JAVED (Member, IEEE) received the master's degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, Pakistan. He has worked with the National Cybercrimes and Forensics Laboratory, Air University, Islamabad, where he is currently a Lecturer with the Department of Cyber Security. He is a member of ACM. He is also a Cybersecurity Researcher and a Practitioner with industry and academic experience. He has reviewed over 150 scientific research articles for various well-known journals. His current research interests include 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 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 a Moderator in the 1st IEEE International Conference on Cyber Warfare and Security (ICCWS). He has authored over 60 peer-reviewed research articles and is supervising/co-supervising several graduate (B.S. and M.S.) students on topics related to health informatics, cybersecurity, mobile computing, and digital forensics. He aims to contribute to interdisciplinary research in computer science and human-related disciplines.

AMANULLAH YASIN received the master's degree in knowledge extraction from data and the Ph.D. degree in data mining and machine learning from the University of Nantes, France. He was a Researcher with the DUKE (Data User KnowledgE) Research Laboratory, School of Engineering of Nantes.

ZUNERA JALIL received the master's degree in computer science with a scholarship from the Higher Education Commission of Pakistan, in 2007, and the Ph.D. degree in computer science with a specialization in information security from the FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2010. She is currently an Assistant Professor with the Department of Cyber Security, Faculty of Computing & Artificial Intelligence, Air University, Islamabad, and a Senior Researcher with the National Cybercrimes and Forensics Laboaratory, National Center for Cyber Security, Islamabad. She has worked as a full-time Faculty Member of 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.

NATALIA KRYVINSKA received the Ph.D. degree in electrical & IT engineering from the Vienna University of Technology, Austria, and a Docent (Habilitation) degree in management information systems from Comenius University in Bratislava, Slovakia. She is currently a Full Professor and the Head of Information Systems Department, Faculty of Management, Comenius University in Bratislava. Previously, she worked as a University Lecturer and a Senior Researcher with the e-Business Department, University of Vienna's School of Business Economics and Statistics. 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.

USMAN TARIQ received the Ph.D. degree in information and communication technology in computer science from Ajou University, South Korea. He is currently a Skilled Research Engineer with Ajou University. His strong background is in *ad-hoc* networks and network communications. He is experienced in managing and developing projects from conception to completion. He has worked on a large international scale and long-term projects with multinational organizations. He is also attached to the College of Computer Engineering and Science, Prince Sattam bin Abdul-Aziz University. His research interests include span networking, security fields, several network security problems, such as botnets, denial-of-service attacks, IP spoofing, and methodologies for conducting security.