IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 21, 2020, accepted June 27, 2020, date of publication June 30, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006082

Hybrid Computerized Method for Environmental Sound Classification

SILVIA LIBERATA ULLO[1], (Senior Member, IEEE), SMITH K. KHARE[2], VARUN BAJAJ[2], (Senior Member, IEEE), AND G. R. SINHA[3], (Senior Member, IEEE)

[1]Department of Engineering, Universita Degli Studi Del Sannio, 821000 Benevento, Italy (e-mail: ullo@unisannio.it).
[2]Department of Electronics and Communication, PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482005, India (e-mail: smith7khare@gmail.com; varunb@iiitdmj.ac.in).
[3]Myanmar Institute of Information Technology (MIIT), Mandalay 05053, Myanmar (e-mail: gr_sinha@miit.edu.mm).

Corresponding author: G. R. Sinha (gr_sinha@miit.edu.mm)

ABSTRACT Classification of environmental sounds plays a key role in security, investigation, robotics since the study of the sounds present in a specific environment can allow to get significant insights. Lack of standardized methods for an automatic and effective environmental sound classification (ESC) creates a need to be urgently satisfied. As a response to this limitation, in this paper, a hybrid model for automatic and accurate classification of environmental sounds is proposed. Optimum allocation sampling (OAS) is used to elicit the informative samples from each class. The representative samples obtained by OAS are turned into the spectrogram containing their time-frequency-amplitude representation by using a short-time Fourier transform (STFT). The spectrogram is then given as an input to pre-trained AlexNet and Visual Geometry Group (VGG)-16 networks. Multiple deep features are extracted using the pre-trained networks and classified by using multiple classification techniques namely decision tree (fine, medium, coarse kernel), k-nearest neighbor (fine, medium, cosine, cubic, coarse and weighted kernel), support vector machine, linear discriminant analysis, bagged tree and softmax classifiers. The ESC-10, a ten-class environmental sound dataset, is used for the evaluation of the methodology. An accuracy of 90.1%, 95.8%, 94.7%, 87.9%, 95.6%, and 92.4% is obtained with a decision tree, k-neared neighbor, support vector machine, linear discriminant analysis, bagged tree and softmax classifier respectively. The proposed method proved to be robust, effective, and promising in comparison with other existing state-of-the-art techniques, using the same dataset.

INDEX TERMS Environmental sound classification, optimal allocation sampling, spectrogram, convolutional neural network, classification techniques.

I. INTRODUCTION

Environment sound is due to numerous sources present in the environment, such as living beings, non-living objects and artificial entities created by humans. These sources contribute in the environment sound, which may be audible as well as non-audible to human ears. The sounds are captured mostly by acoustic sensors, radar systems [1], [2] and subjected to further processing in various sound analysis and applications. The environment sounds, generated by various living beings and non-living objects, needs to be classified in certain categories in order to be used for different purposes, such as security, crime investigation, automated operation of robotic-like vehicles, weather forecasting, environment monitoring [3], etc. Current literature reports numerous studies sand research contributions in the area of environment sound classification (ESC). One such important work on ESC helps detecting natural sounds of environment successfully [3], by using optimum allocation sampling and employing various support vector machine (SVM) and extreme machine learning based classifiers, which operate on OAS-EMD features. Through the sampling method adopted in [3], [4], important performance parameters were studied and determined, which are very useful in the classification of the sounds. In [5], three short audio clips of environment sounds were classified using convolution neural network (CNN) having two convolution layers and max-pooling scheme. This work needed a long training time despite the limited number of datasets used to train the network. A four convolution layer based deep learning network has been used in [6], that utilized Dempster-Shafer evidence theory to produce classification. In this latter case, several

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Li.

CNN based classification methods were outperformed and the variation in classification accuracy for varying sound types, such as car horn, and dog barking, was reported. As we know the sound may be due to various sources [2], [7], [8], and Fisher discriminant model and Gaussian mixture model can help in classifying sounds when a mixture of them is present in the environment [7], by producing a generalized classification. Performance measures, such as spectrogram coefficients and scalar features, have been used for sound classification in [9], where ordinary neural network were employed and a large variety of environment sounds were classified. In another work based on CNN, microcontroller and some other hardware were used [10], with main emphasis on noise characterization of environment sounds in addition to the classification, and quantization issues were major concerns in this work. A computational model to detect the environmental auditory tones temporal deviancy and the frequency saliency has been proposed in [11]. Classification of environmental sound has been accomplished with a filter bank and a Hidden Markov model in [12]. A filter bank and a dimensional reduction are used for signal conditioning while the Hidden Markov model is used to classify the environmental sound. Identification of environmental sound has been accomplished by the means of spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), and Cross Recurrence Plot (CRP) in [13]. Framing and smoothing techniques based on non-overlapping and Hanning window are proposed in [14], where smoothed environmental signals have been used to extract several statistical measures which have been classified by using different classification techniques. Classification of environmental sound has been done by filtering their Fourier transform with a Hanning window for the extraction of the features. These features have been classified by using different deep classififcation methods in [15]. Other literature on ESC refers to [16]–[20], where different types of methods without any robust approach and specific objective are presented. CNN has been also used in [21] for classification of environment sound with augmented training set, towards finding the impact of the classification with and without the augmentation sets. Max-pooling and CNN as used in [5], has been enhanced for sound classification that provided dilation rate for more number of sounds, in addition to evaluation of classification accuracy. However, dilated CNN affects the accuracy of classification to significant extent that appears reduced in this work.

The methods proposed in this literature are limited by the usage of feature extraction methods. Moreover, the majority of the methods have been limited by their performance. Lack of standardized methods and limited performance create an immediate need for new proposals among the scope of environmental sound identification. Also, feature extraction and classification of environmental sound require huge statistical analysis. This motivates us to present a simple, robust, and effective method for the classification of environmental sound. The proposed methodology uses optimal allocation sampling (OAS) to reduce the dimensionality and obtain representative signals from the different classes. The data obtained from OAS are given as input to a short-time Fourier transform (STFT), and the spectrograms obtained from the STFT as input to convolutional neural networks (CNNs). The deep features obtained from the CNNs are classified by using different classification techniques. The contribution of the proposed methodology is summarized as follows:

- Exploring the detailed analysis of the environmental sound classification dataset.
- Reduction of the dimensionality of data by using optimum allocation sampling.
- Transforming the time-domain signals into time-frequency-amplitude representation using a short-time Fourier transform.
- Automatic extraction of various deep features from spectrograms using different CNNs.
- Analysis of different classification techniques to classify different classes of environmental sound.
- Testing effectiveness of the proposed methodology by comparing it with existing state-of-the-art.

The remainder of the paper is organized as follows: Section II describes the methodology, results are presented in Section III, discussion of the proposed method with regards to the existing state-of-the-art is presented in Section IV and finally, in Section V conclusions are given.

II. METHODOLOGY

The proposed methodology presents an effective and robust method for the accurate classification of environmental signals. The methodology is composed of an environmental sound dataset, optimum allocation sampling, short-time Fourier transform, convolutional neural networks, and classification techniques. The stepwise implementation of the proposed work is shown in FIGURE 1.

A. DATA-SET

The methodology uses a free sound database ESC-10 of field recordings. The details of the dataset are available in [22]. The dataset contains 10 classes of different environmental sound annotated as classes (C). The 10 classes have been classified into three general groups of sounds represented by

- transient/percussive sounds, sometimes with very meaningful temporal patterns (sneezing (C-1), clock ticking (C-2), dog barking (C-3)),
- more or less structured noise/soundscapes (sea waves (C-4), fire crackling (C-5), rain (C-6), chainsaw (C-7), helicopter (C-8)),
- sound events with strong harmonic content (crowing rooster (C-10), crying baby (C-9)).

Each signal contains a 5-second-long recordings of audio events (shorter events were padded with silence as needed). The extracted samples were reconverted to a unified format (44.1 kHz, single-channel, Ogg Vorbis compression at 192 kbit/s). Each class consists of forty signals with a total

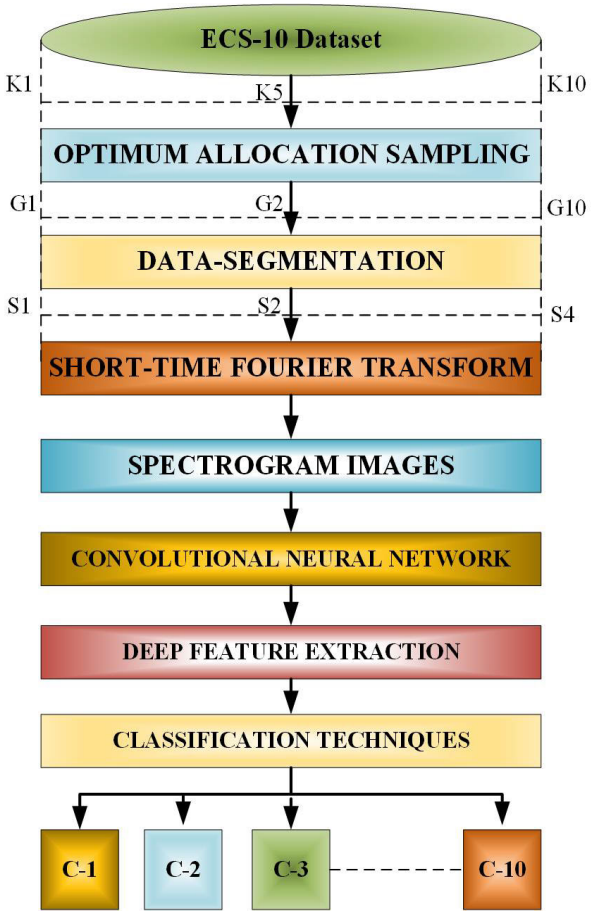

FIGURE 1. Stepwise flow of the proposed method.

length of 160704 samples in each signal. The sound signals of each class are shown in FIGURE 2.

B. OPTIMUM ALLOCATION SAMPLING

The number of samples in each signal has higher dimensionality. Getting insight information from such high dimensional data is very difficult. To overcome this, optimum allocation sampling is used (OAS). OAS reduces the dimensionality of the signal by retaining properties in a signal.

1) DATA PARTITION

Most of the signals existing in real-time exhibit non-stationarity and non-linearity. A long-duration signal exhibiting a non-linearity and non-stationarity may show stationarity and linearity if partitioned into smaller sections. Analysis and processing of stationary signals provide significant insight information. Motivated by this, a number of signals in each class of environmental sound are partitioned into smaller sections. Partition is performed with respect to a specific period. The partitioned section of each signal is called *groups* denoted by $K_1, K_2, .., K_i$, and each partition consists of number of observations denoted by $G_1, G_2, \ldots, G_i$. Every signal in each class must be non-overlapping.

2) OPTIMUM ALLOCATED SAMPLE SIZE

This section aims to get the most representative signals from the group of signals. The most informative samples are selected by considering the groups with minimum variance [23], [24]. The sample size of a group is large if the variability within a group is large. If the variability within a group is small then the sample size is also small. Moreover, OAS is used to find a total sample size of the entire dataset (denoted by p) that needs to be allocated among m groups with the smallest variability. The best sample size is known as the optimum allocated sample for the i^{th} group and it is represented by

$$p_i = \frac{G_i\sqrt{\sum_{l=1}^{L}(V_{il})^2}}{\sum_{i=1}^{m} G_i\sqrt{\sum_{l=1}^{L}(V_{il})^2}} \times p \tag{1}$$

where the size of the i^{th} group is denoted by G_i; the total number of samples in the stratification process is p and V_{il} is the standard deviation of l^{th} signal in the i^{th} group. The total samples of the stratification process are represented by

$$p = \frac{p_0}{1 + \frac{p_0 - 1}{PS}}$$
$$p_0 = \frac{z^2 \times x \times y}{d^2} \tag{2}$$

where p_0 is the initial sample size; and PS is the population size; z (Z-value) is the standard normal variate; d is the desired level of precision; x is a particular characteristics in the dataset and $y = 1\text{-}x$. The total sample size (p) is evaluated by a sample size calculator using a survey software [25]. The representative signals with reduced dimensions obtained by using OAS for different classes are shown in FIGURE 3. The number of samples in each signal is obtained as 16020 respectively.

C. SHORT-TIME FOURIER TRANSFORM

A signal in one domain may not provide insight information directly. To capture the detailed insight of the signal, it may be necessary to transform this latter into another domain. To study the signal in a representative form, it is required to study it in the time-frequency domain simultaneously. The spectral variations in time-frequency-amplitude are simultaneously captured by time-frequency representation. Short-time Fourier transform (STFT) is one such method which is an advanced version of the Fourier transform that provides temporal details of signals in both the domains. Through the STFT, a signal is partitioned in a fixed-sized time-domain signal using a window. A partitioned part is taken off and Fourier transform is applied to study various properties of the signal. In other words, STFT is evenly spaced by using identical and symmetric bandpass filters in the frequency domain. The mathematical formulation of any signal $s(t)$ is represented by [26]

$$S(f, t) = \int_{-T}^{T} s(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau \tag{3}$$

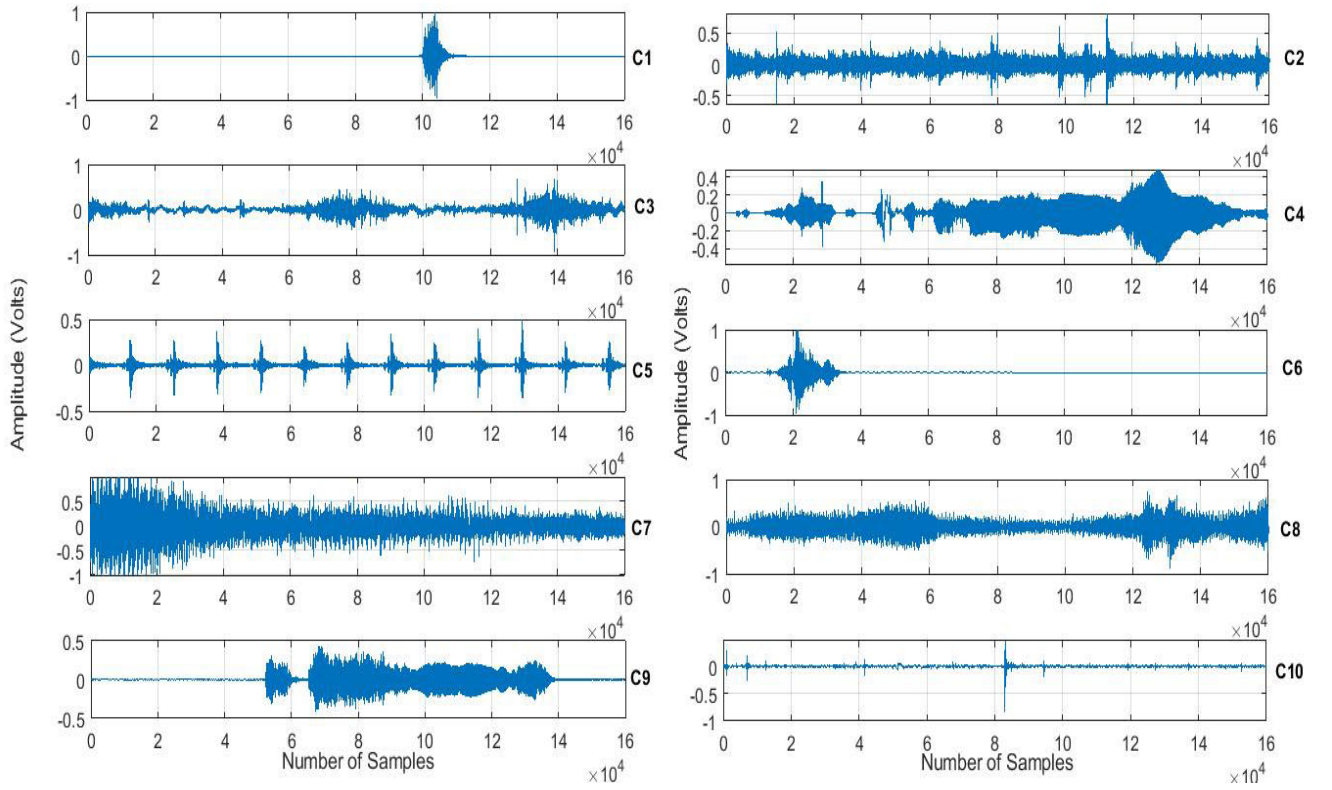

FIGURE 2. Original environmental signals.

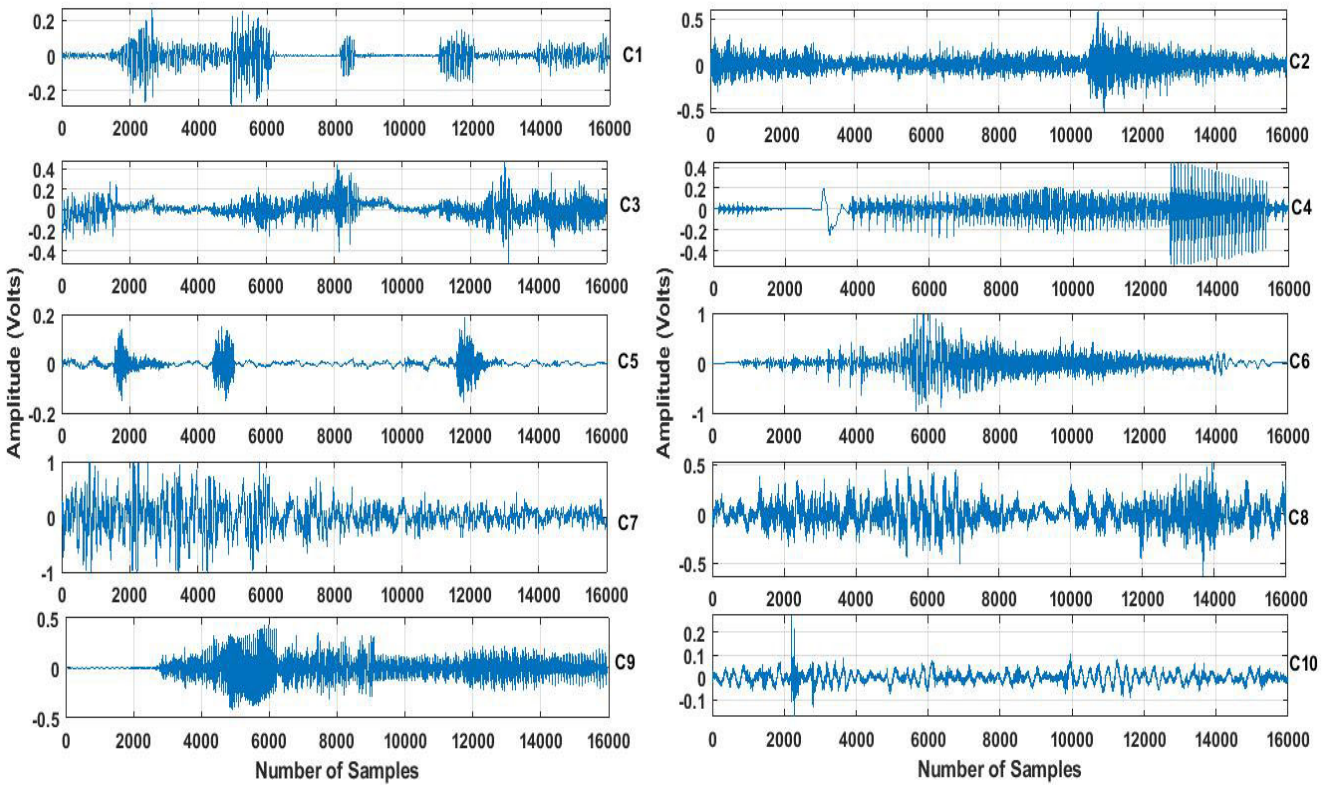

FIGURE 3. Reduced environmental signals using OAS.

where $w(t)$ is the windowing function. The signal $s(t)$ is assumed to be stationary inside the window duration. The type and length of the window must be the same and equal for all the partitioned segments of the signal. The magnitude squared value of the time-frequency representation obtained by STFT is the spectrogram represented

by

$$Spectrogram = |S(f, t)|^2 \tag{4}$$

D. DEEP FEATURE EXTRACTION

Conventional methods require a lot of qualitative and quantitative analysis for the extraction of features. Moreover, parameter tuning is another issue with the conventional feature extraction method. To overcome this, deep feature extraction is employed with the aid of a convolutional neural network (CNN). CNN is a sub-domain of machine learning that comprises of input, hidden, and output layer. The input layer takes images as input for the automatic extraction of numerous features. The hidden layer consists of a convolutional layer, max-pooling layer, and batch normalization operation. The classification task is carried out by the output layer [27], [28]. In this paper, the extraction of deep features are accomplished by CNN, and classification is done by different external classifiers. A hidden block is the driving power of CNN which is responsible for the extraction of deep features. The hidden layer embodies the convolutional layer, dropout, pooling layer, and batch normalization. The function of each block in a hidden layer is explained below:

- **Convolution Layer:** It is a set of filters with learnable parameters. The learnable parameters of filters are tuned automatically with the advancement in training. The size (height and width) of filters is smaller than that of an input image. An input volume is convolved with each filter to evaluate activation maps. The convolution between input and filters is computed at every position by sliding the filters across the height and width of the image. The output of a convolution are 2-dimensional feature maps. The feature maps produced by convolution are followed by Rectified Linear Unit (ReLu) to increase the non-linearity.
- **Batch Normalization:** Batch normalization normalizes the inputs by calculating the mean and variance over each input channel.
- **Pooling Layer:** It is another block of the hidden layer. It is used to reduce the number of parameters and computation by reducing the spatial size. The amount of parameters and computation is reduced with the help of down-sampling by merging the features with various operators. The down-sampling operations allow going deeper into the network.
- **Dropout Layer:** Dropout is the last block of the hidden layer which is applied either on the part or all the features. Dropout helps the network to prevent overfitting.

The above layers are used to construct any CNN. The CNN architecture for deep feature extraction is shown in FIGURE 4. There are several networks available that are used to extract deep features. These networks consist of a fixed number of hidden layers. In this work, two well-known pre-trained CNNs namely AlexNet and VGG (Visual Geometry Group)-16 are employed for automatic feature extraction. The details of the networks are available in [29], [30].

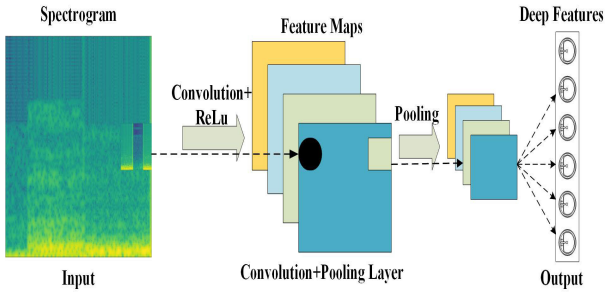

FIGURE 4. Architecture of CNN for deep feature extraction.

The detailed discussion about the feature extraction from images using CNNs is explained in [31].

E. CLASSIFICATION TECHNIQUES

The deep features extracted from the CNNs are classified by using different types of classification techniques. Classification algorithms are used to segregate two or more classes by certain mathematical principles. In this paper, multiple classification techniques are employed for the classification of environmental sound signals. Six types of classification algorithms are used namely decision tree, k-nearest neighbor (k-NN), support vector machines (SVM), linear discriminant analysis (LDA), bagged tree (BT), and softmax. Three kernels are employed with decision tree variants (fine, medium, and coarse), six kernels of kNN variant (fine, medium, coarse, cosine, cubic, and weighted). The reason for using multiple classification techniques is due to the 'No Free Lunch Theorem'. The performance of one classification technique can be overthrown by another [32]. The details of the classification methods and their kernels are available in [33]–[37].

III. RESULTS

The proposed method uses hybrid structure combing OAS, STFT, CNN and classification techniques to classify different class of environmental sounds. ESC-10 dataset is employed for the evaluation of the proposed methodology. To maintain the effectiveness, common experimental platform is maintained throughout the methodology. The experiments have been carried out on MATLAB (2018R). A computer with 8 GB RAM, intel i7 third generation processor of 3.4 GHz, 64 bit memory has been used. In order to reduce the dimensionality of the dataset, OAS is applied. Each signal from every class is partitioned into ten equal segments ($i = 10$) such that $G = G_1 + G_2 + \ldots + G_i$ and $G_i = 16070$. The Z value is obtained as 1.28 when computed from [25]. The confidence level is taken as 99%, estimator x is 50%, and $d = 0.001$ is considered. A total of g_i is 16020 obtained such that $g = g_1 + g_2 + \ldots + g_i$.

TABLE 1 represents the optimum sampling allocated values of each signal for all the ten classes. The signals of the optimum sample size are used for further computation. Before applying the STFT, the data of every signal are split into an equal sample size of 4000. As stated earlier, each

TABLE 1. Accuracy of different CNN features with various classifiers.

Groups	Size (g_i, G_i)	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10
K-1	g_1	5056	1625	2514	1231	4486	526	2487	3624	1424	958
	G_1	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-2	g_2	421	1289	584	1	198	0	0	0	1	0
	G_2	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-3	g_1	0	214	3214	4532	4857	5214	1524	1265	3574	3654
	G_1	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-4	g_3	568	0	486	2	0	231	1	3	0	1325
	G_3	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-5	g_4	5124	12	562	1412	0	0	117	124	0	4
	G_4	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-6	g_5	265	1658	365	7	0	0	2	72	0	0
	G_5	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-7	g_6	3256	2	1214	2314	3214	0	3265	1245	1265	189
	G_7	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-8	g_8	0	6541	1245	0	0	632	0	26	0	2415
	G_8	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-9	g_9	125	2314	1211	0	0	770	0	3	98	1221
	G_9	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
K-10	g_{10}	1205	2365	4625	6521	3265	8647	8624	9658	9658	6254
	G_{10}	16070	16070	16070	16070	16070	16070	16070	16070	16070	16070
OAS	$Total(g)$	16020	16020	16020	16020	16020	16020	16020	16020	16020	16020
	$Total(G)$	160700	160700	160700	160700	160700	160700	160700	160700	160700	160700

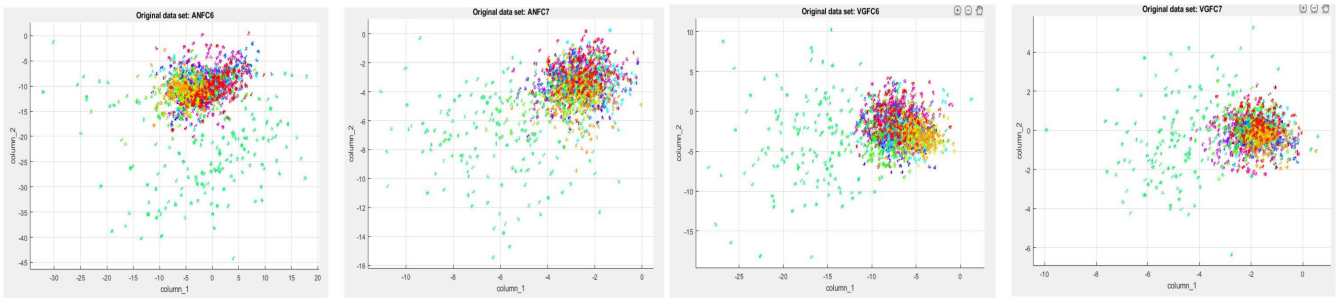

FIGURE 5. Scatter plot of features obtained by DCNN.

class consists of a total of 40 signals. The split signal of each class becomes 160. A non-overlapping Hanning window of size 120 is used. The spectrograms obtained from STFT are given as an input to AlexNet and VGG-16. The pre-trained AlexNet and VGG-16 take images of size 224 × 224 and 227 × 227. AlexNet and VGG-16 consist of 5 and 13 convolutional layers respectively. The feature maps obtained after convolution operations are 2-dimensional. To convert the feature-maps into a single dimension, a fully connected layer is used. In this paper, fully connected layers (FC-6 and FC-7) are used. The scatter plots features obtained by AlexNet and VGG-16 with FC-6 and FC-7 are shown in FIGURE 5. As evident from the figure, all the feature-maps are different and distinguishable. Each class of environmental sound contains a total of 4096 features. Thus, a total feature matrix is obtained by AlexNet and VGG-16 with FC-6 and FC-7 of dimensions 4096 × 160 for each class. This matrix is given as input to different classifiers. The proposed methodology uses 10-cross validation. In this, the input feature matrix is partitioned into 10 disjoint subsets randomly with 9 subsets used for training and the remaining subset for testing. Training and testing are iterated 10 times such that each data point in the feature-matrix is utilized effectively. Six types of classifiers are employed for the classification: decision tree (fine, medium, and coarse kernels), k-NN (fine, medium, coarse, cosine, cubic, and weighted kernels), LDA, SVM, BT, and softmax (SM).

The parameter tuning is maintained uniform for each variant of a classifier, for all the kernels. The number of splits is kept at 20 for the decision tree, decision metric, decision weight, and a number of neighbors is set to Euclidian, squared inverse, and 2 for k-NN. Full covariance structure is used in the case of LDA, box constraint level is set to 1, automatic kernel scale mode and one v/s all multiclass method are used for SVM. For BT, decision tree learner is used, a number of learners and the learning rate is set to 20 and 0.001,

TABLE 2. Accuracy of different CNN features with various classifiers.

Feature Extractor	Classifier Variant	Kernel	Parameter Setting	Accuracy
AlexNet (FC-6)	Decision Tree	Fine	Number of splits 20	89.9
		Medium		89.5
		Coarse		47.9
AlexNet (FC-7)	Decision Tree	Fine	Number of splits 20	89.2
		Medium		89.8
		Coarse		48.4
VGG-16 (FC-6)	Decision Tree	Fine	Number of splits 20	88.4
		Medium		88.1
		Coarse		48
VGG-16 (FC-7)	Decision Tree	Fine	Number of splits 20	90.1
		Medium		89.8
		Coarse		48.3
AlexNet (FC-6)	k-NN	Fine	Distance Metric = Euclidian Distance Weight = Squared inverse No of neighbors = 2	94.5
		Medium		92.5
		Coarse		74.6
		Cosine		84.3
		Cubic		89.6
		Weighted		93.6
AlexNet (FC-7)	k-NN	Fine	Distance Metric = Euclidian Distance Weight = Squared inverse No of neighbors = 2	94.6
		Medium		93.6
		Coarse		83.1
		Cosine		84.4
		Cubic		92.4
		Weighted		94.1
VGG-16 (FC-6)	k-NN	Fine	Distance Metric = Euclidian Distance Weight = Squared inverse No of neighbors = 2	95.8
		Medium		92.9
		Coarse		73.3
		Cosine		83.1
		Cubic		90.3
		Weighted		93.9
VGG-16 (FC-7)	k-NN	Fine	Distance Metric = Euclidian Distance Weight = Squared inverse No of neighbors = 2	95.4
		Medium		94.2
		Coarse		74.8
		Cosine		83.8
		Cubic		89.6
		Weighted		94.9
AlexNet (FC-6)	LDA	–	Covariance Structure = Full	91.1
AlexNet (FC-7)				92.3
VGG-16 (FC-6)				94.1
VGG-16 (FC-7)				94.7
AlexNet (FC-6)	SVM	Gaussian	Box-constraint level = 1 Kernel scale mode = Automatic Multiclass method = One v/s All	75.3
AlexNet (FC-7)				81.4
VGG-16 (FC-6)				85.6
VGG-16 (FC-7)				87.6
AlexNet (FC-6)	BT	–	Learner Type = Decision Tree Maximum Number of Splits = 1599 Number of learners = 20 Learning Rate = 0.001	93.8
AlexNet (FC-7)				95.3
VGG-16 (FC-6)				95.6
VGG-16 (FC-7)				95.4
AlexNet (FC-6)	SoftMax	–		90.8
AlexNet (FC-7)				91.3
VGG-16 (FC-6)				91.6
VGG-16 (FC-7)				92.4

and a maximum number of splits is 1599. The classification accuracy obtained for AlexNet and VGG-16 with FC-6 and FC-7 is shown in TABLE 2.

The maximum accuracy obtained for AlexNet (FC-6) for fine kernel using a decision tree is 89.9%. An accuracy of 89.8% is highest for AlexNet (FC-6) using a medium (M) kernel of a decision tree. The highest classification accuracy for VGG-16 (FC-6) and VGG-16 (FC-7) is obtained for a fine kernel of a medium tree having value 88.1% and 90.1% respectively. With k-NN, the highest accuracy

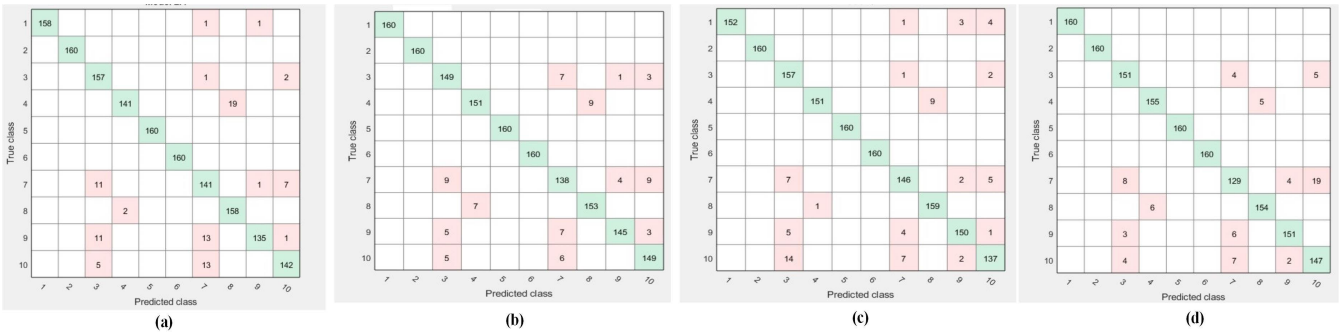

FIGURE 6. Confusion matrix.

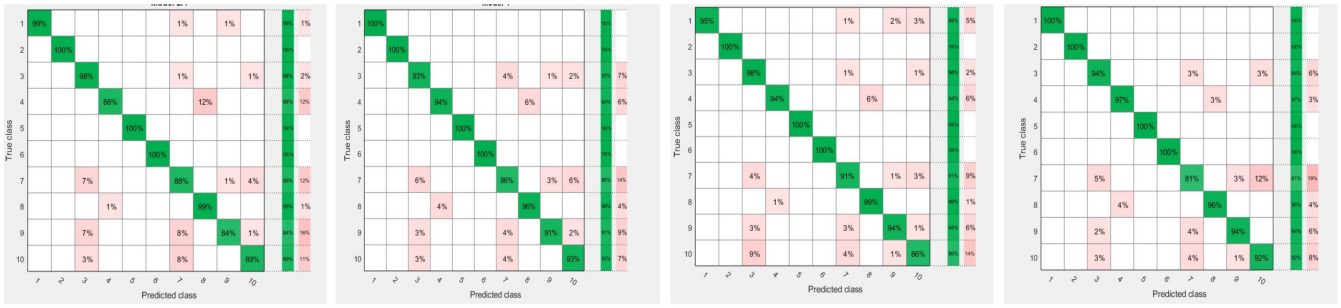

FIGURE 7. False negative rate.

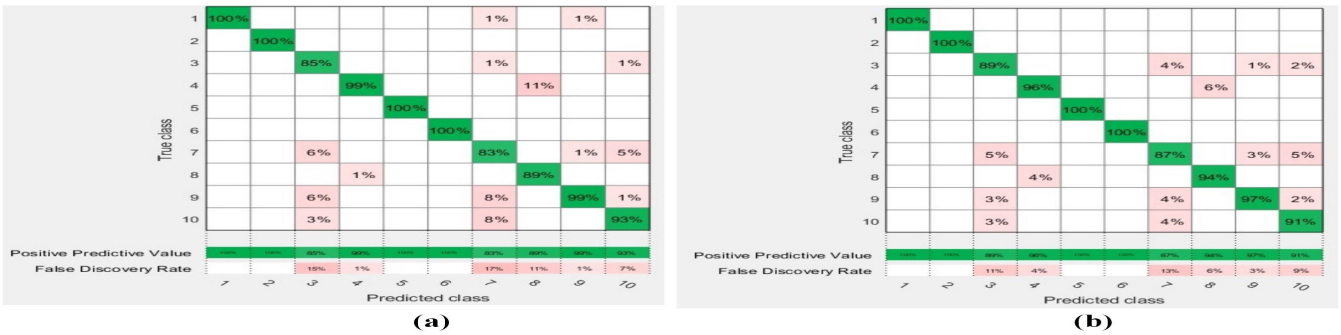

FIGURE 8. False discovery rate of AlexNet with FC-6 and FC-7.

obtained for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7) is with fine (F) kernel having an accuracy of 94.5%, 94.6%, 95.8%, and 95.4%, respectively. Accuracy obtained with LDA for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7) is 91.1%, 92.3%, 94.1%, and 94.7%, respectively. SVM provides an accuracy of 75.3%, 81.4%, 85.6%, and 87.6% for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7). Classification accuracy of 93.8%, 95.3%, 95.6%, and 95.4% is obtained with BT for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7) while with softmax classifier an accuracy of 90.8%, 91.3%, 91.6%, and 92.4% is obtained for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7). As seen from the TABLE 2, a fine kernel is the best performer and coarse kernel shows the worst performance for decision tree and k-NN. The best separation of all the classes is provided by k-NN for AlexNet (FC-6) and VGG-16 (FC-6), while for AlexNet (FC-7) bagged is superior and there is a tie between k-NN and bagged tree for VGG-16 (FC-7) features.

The confusion matrix of best performing classifiers is shown in FIGURE 6, where (a), (b), (c), and (d) shows confusion matrix for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7) with k-NN and BT classifiers. For AlexNet (FC-6) and VGG-16 (FC-6), C-2, C-5, and C-6 provide perfect classification. For AlexNet (FC-7) and VGG-16 (FC-7), C-1, C-2, C-5, and C-6 provides the best results. The false-negative rate for each class of environmental sound is shown in FIGURE 7, where (a), (b), (c), and (d) show the false-negative rate of each class for AlexNet (FC-6), AlexNet (FC-7), VGG-16 (FC-6), and VGG-16 (FC-7). The values in green show the true posi-

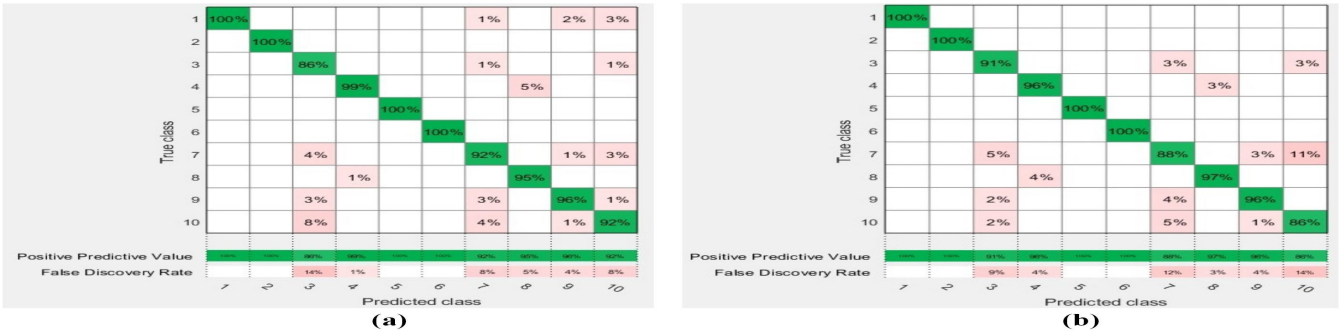

FIGURE 9. False discovery rate of VGG-16 with FC-6 and FC-7.

TABLE 3. Comparison of accuracy with existing method using same dataset.

Authors	Methods	Classifiers	Accuracy
Piczak [5]	Baseline Machine	k-NN	66.7%
		SVM	67.5%
		RFE	72.7%
Pillos et. al [14]	ZCR, MFCC,	RFE	73.75%
	SF & SC	MLP	74.5%
Boddapati et. al [13]	Spectrogram	AlexNet	86%
		GoogleNet	86%
Ahmad et. al [3]	EMD	ELM	77.61%
		LS-SVM	87.25%
Hertel et. al [15]	DNCN		77.1%
			83.7%
Pareta et. al [4]	OAS	MC-LSSVM	85.43%
Proposed	OAS, STFT,	FT	90.1%
	& CNN	MT	89.8%
		F-kNN	95.8%
		M-kNN	94.2%
		C-kNN	83.1%
		Co-kNN	84.4%
		Cu-kNN	92.4%
		W-kNN	94.9%
		LDA	94.7%
		SVM	87.9%
		BT	95.6%
		SM	92.4%

tive rate, while the values in red give a false-negative rate. The false discovery rate of AlexNet (FC-6) and AlexNet (FC-7) is shown in FIGURE 8 (a) and (b). The false discovery rate of VGG-16 (FC-6), and VGG-16 (FC-7) is shown in FIGURE 9 (a) and (b). The values in green give details of positive prediction value and values in red denote the false discovery rate.

IV. DISCUSSION

The effectiveness of the proposed methodology is tested by comparing the classification accuracy with the existing state-of-the-art using the same dataset. TABLE 3 presents the accuracy comparison with different methodologies. Piczak in [5] used a baseline machine method to extract different information from the sound signals. The features obtained by using the baseline machine is classified by using k-NN, SVM, and random forest ensemble (RFE) classifiers. The model in [5] provided an accuracy of 66.7%, 67.5%, and 72.7% with k-NN, SVM, and RFE. The method proposed by Pillos et. al. in [14] evaluated different feature extraction techniques. The features extracted by using zero-crossing rate (ZCR), Mel-frequency cepstral coefficient (MFCC), spectral flatness (SF), and spectral centroid (SC) have been classified by using a multilayer perceptron (MLP) and RFE models. The method in [14] managed to achieve an accuracy of 73.75% and 74.5% with RFE and MLP. Boddapati et. al. in [13] used spectrogram method to classify the signals using AlexNet and Google net. The method proposed in [13] achieved an accuracy of 86%. In [3], Ahmad et. al. used empirical mode decomposition (EMD) for the extraction of several features. These features have been classified by using extreme learning machine (ELM) and least square support vector machine (LS-SVM). The method proposed by Ahmad et. al. achieved an accuracy of 77.7% and 87.25% with ELM and LS-SVM, respectively. The method in [15] by Hertel et. al. used a deep neural convolution network (DNCN) for feature extraction and classification. This model managed an accuracy of 77.1% and 83.7%, respectively. Feature extracted by OAS has been classified using multiclass LSSVM (MC-LSSVM) in [4] by Pareta et. al. and this model achieved an accuracy of 85.43%.

The methods proposed until now by the researchers have been limited in terms of performance. Hence an effective and robust method is required to classify environmental signals accurately. In the present work, authors aim to propose a method in which the dimension of data is reduced by OAS. The reduced data are then used to be transformed into images by STFT. Several features have been extracted from the spectrograms by using two pre-trained CNNs. These features are classified by using different classification techniques. An accuracy of 90.1% and 89.8% is achieved with fine (FT) and medium tree (MT). Accuracy obtained with fine (F), medium (M), coarse (C), cosine (Co), cubic (Cu), and weighted (W) kernels of k-NN is 95.8%, 94.2%, 83.1%,

84.4%, 92.4%, and 94.9%, respectively. The accuracy of 94.7%, 87.9%, 95.6%, and 92.4% is achieved with LDA, SVM, BT, and SM classifiers, respectively. The results of the proposed method seem to be promising and well ahead from the performance of existing methodologies.

V. CONCLUSION

The environmental sound provides lot of information that can be used in various fields. Accurate identification of environmental sound is required for modeling any system. Environmental signals are non-stationary. For this, an adaptive, robust, and effective methodology is needed for correct classification of environmental signals. To this concern, a hybrid model combing OAS, STFT, CNN, and classification technique is proposed. The dimensionality reduction is performed by OAS that reduces the burden of computation, as well. Time-frequency information is captured simultaneously using STFT. CNNs are used to extract deep features and are classified by using different classifiers. The proposed method provides an accuracy of 95.8% with a fine kernel of k-nearest neighbor. The comparison shows that the proposed method provides significant improvement in the separation of environmental signals by about 9%. Thus, the proposed method proved to be promising and can be used to model a real-time environmental sound detection of natural sounds.

REFERENCES

[1] P. Addabbo, M. di Bisceglie, C. Galdi, and S. L. Ullo, "The hyperspectral unmixing of trace-gases from ESA SCIAMACHY reflectance data," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2130–2134, Oct. 2015.

[2] P. Addabbo, C. Clemente, and S. L. Ullo, "Fourier independent component analysis of radar micro-Doppler features," in *Proc. IEEE Int. Workshop Metrol. Aerosp. (MetroAeroSpace)*, Jun. 2017, pp. 45–49.

[3] S. Ahmad, S. Agrawal, S. Joshi, S. Taran, V. Bajaj, F. Demir, and A. Sengur, "Environmental sound classification using optimum allocation sampling based empirical mode decomposition," *Phys. A, Stat. Mech. Appl.*, vol. 537, Jan. 2020, Art. no. 122613.

[4] A. Pareta, S. Taran, V. Bajaj, and A. Sengur, "Automatic environment sounds classification using optimum allocation sampling," in *Proc. 4th Int. Conf. Robot. Autom. Eng. (ICRAE)*, Nov. 2019, pp. 69–73.

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[6] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, Apr. 2019.

[7] K. Haddad, W. Song, and X. Valero, "Environmental sound classification in realistic situations," in *Proc. Forum Acusticum*, Kraków, Poland, Sep. 2014, p. 6.

[8] M. Carminati, O. Kanoun, S. L. Ullo, and S. Marcuccio, "Prospects of distributed wireless sensor networks for urban environmental monitoring," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, no. 6, pp. 44–52, Jun. 2019.

[9] S. A. Mitilineos, S. M. Potirakis, N.-A. Tatlas, and M. Rangoussi, "A two-level sound classification platform for environmental monitoring," *J. Sensors*, vol. 2018, pp. 1–13, Jun. 2018.

[10] J. O. Nordby, "Environmental sound classification on microcontrollers using convolutional neural networks," M.S. thesis, Dept. Sci. Technol., Norwegian Univ. Life Sci., Ås, Norway, 2019.

[11] Y. Su, J. Wang, K. Zhang, and K. Madani, "Computational modeling of environment deviant sound detection based on human auditory cognitive mechanism," *Biologically Inspired Cognit. Archit.*, vol. 24, pp. 87–97, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212683X18300215

[12] C. Salazar-Garcia, R. Castro-González, and A. Chacón-Rodriguez, "RISC-V based sound classifier intended for acoustic surveillance in protected natural environments," in *Proc. IEEE 8th Latin Amer. Symp. Circuits Syst. (LASCAS)*, Feb. 2017, pp. 1–4.

[13] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[14] A. Pillos, K. Alghamidi, N. O. Alzamel, V. Pavlov, and S. Machanavajhala, "A real-time environmental sound recognition system for the Android OS," in *Proc. Detection Classification Acoustic Scenes Events*, Budapest, Hungary, Sep. 2016, pp. 1–5.

[15] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3407–3411.

[16] T. Miano, "Hear and see: End-to-end sound classification and visualization of classified sounds," *PeerJ PrePrints*, vol. 6, Oct. 2018, Art. no. e27280.

[17] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2721–2725.

[18] L. Ma, D. Smith, and B. Milner, "Environmental noise classification for context-aware applications," in *Database and Expert Systems Applications*, V. Mařík, W. Retschitzegger, and O. Štěpánková, Eds. Berlin, Germany: Springer, 2003, pp. 360–370.

[19] S. Chachada and C.-C.-J. Kuo, "Environmental sound recognition: A survey," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Oct. 2013, pp. 1–9.

[20] R. Catanghal, T. Palaoag, and C. Dayagdag, "Environmental acoustic transformation and feature extraction for machine hearing," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, vol. 482, Mar. 2019, Art. no. 012007.

[21] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[22] K. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[23] S. Taran, V. Bajaj, and S. Siuly, "An optimum allocation sampling based feature extraction scheme for distinguishing seizure and seizure-free EEG signals," *Health Inf. Sci. Syst.*, vol. 5, no. 1, p. 7, Dec. 2017.

[24] S. Siuly, H. Wang, and Y. Zhang, "Detection of motor imagery EEG signals employing Naïve Bayes based learning process," *Measurement*, vol. 86, pp. 148–158, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0263224116001469

[25] *Sample Size Calculator*. Accessed: Apr. 24, 2020. [Online]. Available: http://www.surveysystem.com/sscalc.htm

[26] H. K. Kwok and D. L. Jones, "Improved instantaneous frequency estimation using an adaptive short-time Fourier transform," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2964–2972, Oct. 2000.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[28] A. Ozdemir and K. Polat, "Deep learning applications for hyperspectral imaging: A systematic review," *J. Inst. Electron. Comput.*, vol. 2, pp. 39–56, Jan. 2020.

[29] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. C. V. Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from AlexNet: A comprehensive survey on deep learning approaches," *CoRR*, vol. abs/1803.01164, 2018. [Online]. Available: http://arxiv.org/abs/1803.01164

[30] P. Kemal and K. O. Koc, "Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all," *J. Artif. Intell. Syst.*, vol. 2, no. 1, pp. 80–97, 2020.

[31] S. Ullo, M. Langenkamp, T. Oikarinen, M. Del Rosso, A. Sebastianelli, F. Piccirillo, and S. Sica, "Landslide geohazard assessment with convolutional neural networks using Sentinel-2 imagery data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul./Aug. 2019, pp. 9646–9649.

[32] S. Adam, S.-A. Alexandropoulos, P. Pardalos, and M. Vrahatis, "No free lunch theorem: A review," in *Approximation and Optimization*. Cham, Switzerland: Springer, 2019, pp. 57–82, doi: 10.1007/978-3-030-12767-1_5.

[33] Y.-Y. Song and Y. Lu, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015.

[34] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated Web usage data mining and recommendation system using K-nearest neighbor (KNN) classification method," *Appl. Comput. Informat.*, vol. 12, no. 1, pp. 90–108, Jan. 2016.

[35] M. Awad and R. Khanna, *Support Vector Machines for Classification*. Berkeley, CA, USA: Apress, 2015, pp. 39–66, doi: 10.1007/978-1-4302-5990-9_3.
[36] A. Tharwat, "Linear vs. Quadratic discriminant analysis classifier: A tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, p. 145, 2016.
[37] M. Arican and K. Polat, "Binary particle swarm optimization (BPSO) based channel selection in the EEG signals and its application to speller systems," *J. Artif. Intell. Syst.*, vol. 2, no. 1, pp. 27–37, 2020.

SILVIA LIBERATA ULLO (Senior Member, IEEE) graduated with laude in electronic engineering at the Faculty of Engineering, Federico II University, Naples. She received the M.Sc. degree from the Massachusetts Institute of Technology (MIT) Sloan Business School of Boston, USA, in June 1992. She is a Researcher with the University of Sannio di Benevento, where she teaches signal theory and elaboration, and telecommunication networks, courses for the degree in electronic engineering and the optical and radar remote sensing as Ph.D. course. She has been with Italel, since September 1992. She served as a Chief of some production lines at the Santa Maria Capua Vetere factory (CE), until January 2000. She has authored 72 research papers in reputed journals and conferences. She is part of the Telecommunications and Remote Sensing group and her research interests mainly deal with signal processing, remote sensing, image and satellite data analysis, machine learning applied to satellite data, ESA Copernicus mission, cognitive radars, sensor networks, telecommunications networks, and smart grids. She is an Industry Liaison for the IEEE Italy Joint ComSoc / VTS Chapter. She is a member of the Academic Senate at University of Sannio, and a National Referent for the FIDAPA BPW Italy Sciences and Technologie Task Force. She awarded with the Marisa Bellisario prize from the homonymous foundation, and with the Marisa Bellisario scholarship from Italel SpA company. She wins a public competition and started working at the Center for Data Processing (CED) in the Municipality of Benevento, from January 2000 to January 2004. In February 2004, she won the research contest at the Faculty of Engineering, University of Sannio, Benevento.

SMITH K. KHARE received the B.E and M.Tech. degrees from Nagpur and Mumbai University, in 2012 and 2015, respectively. He was working as an Assistant Professor with the Yeshwantrao Chavan College of Engineering, G.H. Raisoni College of Engineering and Shri Ramdeobaba College of Engineering and Management, Nagpur, India. He is currently a Research Scholar with the PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, India. He has authored/coauthored nine publications in various high impact factor and peer-reviewed, journals. He has published two paper in international conference. His research interests include biomedical signal processing, pattern recognition, machine learning, and deep neural networks. He is serving as a reviewer in IEEE, Elsevier, and several other reputed journals.

VARUN BAJAJ (Senior Member, IEEE) received the B.E. degree in electronics and communication engineering from Rajiv Gandhi Technological University, Bhopal, India, in 2006, the M.Tech. degree (Hons.) in microelectronics and VLSI design from the Shri Govindram Seksaria Institute of Technology and Science, Indore, India, in 2009, and the Ph.D. degree in electrical engineering from the Indian Institute of Technology Indore, India in 2014. He worked as a Visiting Faculty at IIITDM, from September 2013 to March 2014. He served as an Assistant Professor with the Department of Electronics and Instrumentation, Shri Vaishnav Institute of Technology and Science, Indore, India, from 2009 to 2010. He has been working as a faculty in electronics and communication engineering with the Indian Institute of Information Technology, Design and Manufacturing (IIITDM) Jabalpur, India, since 2014. He has edited Modeling and Analysis of Active Biopotential Signals in Healthcare- Volume 1, 2 published in IOP books. He has authored more than 90 research papers in various reputed international journals/conferences like IEEE Transactions, Elsevier, Springer, and IOP. The citation impact of his publications is around 1779 citations, H-index of 19, and i10 index of 40 (Google Scholar May 2020). He has guided three Ph.D. Scholars and five M.Tech. Scholars. His research interests include biomedical signal processing, image processing, time-frequency analysis, and computer-aided medical diagnosis. He was a recipient of various reputed national and international awards. He served as a Subject Editor of *IET Electronics Letters*, from November 2018 to June 2020. He is also serving as a Subject Editor-in-Chief of *IET Electronics Letters.* He is contributing as an active technical reviewer of leading International journals of IEEE, IET, and Elsevier.

G. R. SINHA (Senior Member, IEEE) received the Ph.D. degree. He is an Adjunct Professor with the International Institute of Information Technology Bangalore (IIITB) and currently deputed as a Professor at the Myanmar Institute of Information Technology (MIIT), Mandalay, Myanmar. He is a Visiting Professor (Honorary) with Sri Lanka Technological Campus Colombo for one year, from 2019 to 2020. He has more than 200 research papers, edited books, and books into his credit. He has edited books for reputed International publishers. He has teaching and research experience of 21 years. He has been the Dean of Faculty and an Executive Council Member of CSVTU and currently a member of Senate of MIIT. He has been delivering ACM lectures as a ACM Distinguished Speaker in the field of DSP, since 2017, across the world. His research interests include biometrics, cognitive science, medical image processing, computer vision, outcome based education (OBE), and ICT tools for developing Employability Skills. He is a Fellow of the Institute of Engineers India and a Fellow of IETE, India. He served as a Distinguished IEEE Lecturer in IEEE India council for Bombay section. He was a recipient of many awards and recognitions at national and international level. He has delivered more than 50 Keynote/Invited Talks and Chaired many Technical Sessions in International Conferences across the world. He has eight Ph.D. Scholars, 15 M.Tech. Scholars, and has been Supervising one Ph.D. Scholar. He is active reviewer and editorial member of more than 12 reputed International journals in his research areas, such as IEEE Transactions, Elsevier journals, and Springer journals.