Audio classification using grasshopper-ride optimization algorithm-based support vector machine

Suryabhan Pratap Singh1 | Umesh Chandra Jaiswal2

1Department of Computer Science and Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh, India
2Department of Information Technology and Computer Application, Madan Mohan Malaviya University of Technology, Gorakhpur, Uttar Pradesh, India

Correspondence
Suryabhan Pratap Singh, Madan Mohan Malaviya University of Technology, Gorakhpur 273010, Uttar Pradesh, India.
Email: spsuryabhan@gmail.com, ucj_jaiswal@yahoo.com

Abstract
The accurate and robust detection of the audio has been widely grown as the speech technology in the area of audio forensics, speech recognition, and so on. However, in real time, it is a challenge to deal with the massive data arriving from distributed sources. Thus, the study introduces a method that effectively deals with the data from the distributed sources using the map-reduce framework (MRF). The map and reduce function in MRF aim at feature extraction and audio classification. The robust classification using the proposed grasshopper-ride optimization algorithm-based support vector machine (G-ROA-based SVM) uses the features, such as multiple kernel Mel frequency cepstral coefficients, spectral flux, spectral kurtosis, and delta-amplitude modulation spectrogram. The proposed G-ROA is the integration of ROA and grasshopper optimization algorithm in tuning the optimal weights of SVM and also, the kernel function in SVM is modified using the Gaussian radial basis function, Gaussian kernel, and polynomial kernels. The experimentation of the proposed method is done using two datasets, namely TUT sound event 2017 dataset and ESC dataset. TUT sound event 2017 dataset consists of eight audio recordings from a single acoustic scene. ESC dataset consists of three parts and 252,400 recordings. The analysis reveals that the proposed audio classification acquired the maximal accuracy of 0.96, minimal false alarm rate, and false rejection rate of 0.022 and 0.0119, respectively.

1 | INTRODUCTION

Audio classification differentiates the audio using the emotion, identity, accents, and other parameters of the speakers, which when performed effectively, can contribute the tasks, like translation from speech-to-speech and automatic recognition of the speech [1]. For example, the emotions of the sound are employed for the translation of the utterances spoken in a language to another [2] that facilitates the ability to handle the non-linguistic data and ensures the practical way of translation. Moreover, enhancing the robustness and generalization of speech recognition is based on the accent variations of individual speech [3]. All the aforementioned examples enable the performance of tasks associated with communication in human–machine speech.

Audio is the rich source of information in case of the detection and classification of the events, and also, audio is employed for applications related to the multimedia event detection [4, 5]. The abnormal conditions comprise the natural damages, the audio reveals important information than a video, and these detected acoustic clues stand as complementary data for the automatic detection of the abnormal conditions [6, 7]. In general, an audio summarization requires the relevant and conspicuous acoustic events, which drag the attention of humans and sketches the simple concepts of the entire audio stream [8]. Techniques for processing of speech and audio, like keyword spotting and voice activity detection (VAD) form a section of summarization in segmentation and detection of audio with significant words [9]. The data contaminated by clouds and the attendant shadows is of great significance for many users’ demands, such as target recognition, classification, segmentation, feature extraction, and so on [10]. Most of the audio classification methods utilize the low-dimensional features for classification since the original audio suffers from a high dimensionality problem [11–13]. For the extraction of the signal details, including energy, time, and...
frequency from the original audio data, spectrograms are employed [1, 14, 15]. To minimize the input space dimensions, the feature extraction is progressed using the linear prediction cepstrum coefficient and Mel-frequency cepstral coefficients (MFCCs) [1]. Though they possess complex sound proportions, many details are common in the audio scenes belonging to the same category. However, matching is a serious criterion in discovering and matching the patterns [16]. Features obtained from the audio, like power spectral density, RASTA analysis, and frequency bands generated from the filter bank in addition to the recurrent neural networks and k-nearest neighbour for classification [17]. Hidden Markov models (HMMs) [18] are employed to model the conventional audio features also with the sound event possesses various representations in case of noisy and clean environments [19]. The traditional methods of classification, like nearest neighbour [20], support vector machine (SVM) [21], HMM [22], deep neural network [23], and Gaussian mixture model [24] are employed. At present, deep learning attracts the research in areas of audio processing for which the deep learning architecture, termed as deep belief network (DBN), is used [25–27]. In the parallel computing frameworks, the performance is high because of the processing speed of classification. The map-reduce function is commonly used to increase the processing speed [28]. The visual features of sounds are built starting from the audio file and are taken from images constructed from different spectrograms, a gammatonegram, and a rhythm image [29].

The main aim of this study is to design and develop an audio classification system in the MRF for the reliable detection of an event that is to improve the accuracy. The proposed system uses mapper and reducer functions, wherein pre-processing and feature extraction are carried out in the mapper phase and the classification is done in the reducer phase. The proposed G-ROA is obtained with the integration of ROA [30] and grasshopper optimization algorithm (GOA) [31]. The proposed G-ROA algorithm inherits the advantages of both the algorithms and the proposed algorithm tunes the optimal weights of SVM in such a way that the detection of audio becomes effective.

The major contribution of the research

G-ROA-SVM for classification of the audio: The audio classification is performed using the G-ROA-SVM classifier. Here, the SVM classifier is trained by the proposed G-ROA, which is the integration of ROA and GOA. The G-ROA tuning the optimal weights of SVM and the standard kernel functions in SVM are modified using radial basis function (RBF), Gaussian kernel, and polynomial kernel, which enhance the accuracy of the audio classification.

The rest of the study is organized as: Section 2 reveals the review of the existing methods with the reasonable challenges that insist the need for the new method of audio classification. The proposed method of audio classification is detailed in Section 3, and Section 4 presents the results of the proposed method. Finally, Section 5 concludes the study with the effective proof for audio classification.

2 | MOTIVATION

In this section, the review of the literature is presented with the challenges faced that motivated the work.

2.1 | Literature survey

In this section, the existing methods are reviewed with eight existing works along with the pros and cons of the developed method. Waldekar and Saha [19] developed a fused system framework, and in the method, the significance was regarding the frame-level statistics of the known spectral features, which formed the input to SVM that proved the outperforming nature of the system. The drawback of the method was regarding the binaural data format. Ali and Talha [32] developed a method for VAD based on unsupervised learning, and hence, there was no need for the training data for distinguishing the voice segments from the unvoiced ones. The drawback of the method was that the samples at the end of the audio were not a part of any frame as a result of the insufficient samples. Wang et al. [27] developed a method, hierarchical-diving deep belief network (HDDBN), which was robust even for noisy situations. The drawback of the method was that when the low-level DBNs tune the network for the iterations, the performance was affected. Phan et al. [16] developed a scene classification method using the convolutional neural networks (CNNs) and label-tree embeddings, which reduced the low-level features as likelihoods of metaclasses that made the learning and matching of templates efficient. The drawback of the method was regarding the size of the segment, which should not be shorter. Thus, more details of the signals were required, leading to the unreliable estimation by the random forest classifier using the posterior probabilities at the time of the feature learning using the label tree embedding (LTE). Souli and Lachi [33] performed the audio classification with the SVM and scattering features. The ability associated with the representation of the nonstationary signals enables the ability to discriminate the events, including filters for time and rhythms. Moreover, the classification accuracy of the method was not good in case of the presence of environmental noises but was effective in the case of clean environments. Wu et al. [1] developed an attention-augmented CNN that enhanced the features generated from the frequency bands. The drawback of the method was regarding the higher number of the local frequency segments as this would increase the parameters in the model, affecting the performance of the system. Hong et al. [34] developed a non-negative matrix factorization depending on the feature learning mechanism that worked robustly even under the high environmental and noisy speech. The method was effective when compared with the other feature selection methods and the drawback was the training data samples were not guaranteed to remain independent. Arumugam and Kaliappan [35] developed a feature selection strategy using the modified bacterial foraging optimization algorithm (MBFOA) that was efficient in the case of multimedia applications, but the performance was questionable.
2.2 Challenges

The challenges of the research are revealed below:

- The classification was better, but the error rates remained high during audio classification [32, 35].
- In deep learning [6] there is a need for a large amount of data and also there is a need for signal-specific feature engineering [32, 34, 36, 37].
- The ‘bag-of-frames’ system is superior to the simpler one-point average approach, while evaluating the other three available audio scenes datasets possessing less within-class variability [19, 30, 31].
- Latent Dirichlet allocation and probabilistic latent semantic analysis methods, there is no assumption regarding the generation of document-topic distribution [38].
- Sparse representations of the noise represented as the basis vectors of the transformation matrix appear to be different from the actual signal [4, 6, 26, 39].

3 PROPOSED AUDIO CLASSIFICATION STRATEGY USING THE OPTIMIZATION ENABLED SVM CLASSIFIER

The audio classification seems to be a complex and tedious process as the processing time is large to deal with the big data. Thus, the study uses the MRF for the classification of the audio to reduce the processing time in such a way that the parallel processing of the huge data is enabled. There are three major steps, the first step is pre-processing, feature extraction, and finally, the audio classification. The proposed system uses mapper and reducer functions, wherein pre-processing and feature extraction are carried out in the mapper phase and the classification is done in the reducer phase. The pre-processing helps to remove the background non-voice region from the signal and enables feature extraction. The feature extraction helps to classify the audio. The audio classification is performed using the G-ROA-SVM classifier. The SVM classifier is trained by the proposed G-ROA, which is the integration of ROA and GOA. For audio classification, a trained SVM classifier is employed in which the kernel is replaced with a new kernel function and the training is performed using the proposed algorithm, G-ROA. Finally, the features are progressed using the proposed G-ROA-based SVM. Figure 1 shows the block diagram of the proposed audio classification strategy.

3.1 Map-reduce framework for audio classification

Map-reduce is a framework for processing and generating a huge amount of data and the structure is organized as a single master node and a number of worker nodes. Let us represent the input data as \( S \), which carries the input audio signals and the database \( S \) resembles the big data. Hence, for the effective audio classification, the input database is split as the subsets of data, which is the role of the master node in the MRF. Let the data subsets are denoted as
where \( n \) refers to the total number of the data subset, and the subsets formed from the database is the input to the mapper phase. The individual audio signal from the database is represented as \( y \).

### 3.1.1 Mapper phase

The mappers are denoted as:

\[
M = \{M_1, M_2, \ldots, M_q, \ldots, M_n\} \tag{2}
\]

It is clear from Equations (1) and (2) that the total number of mappers is equal to the total data subsets. The features are extracted from the pre-processed audio signals, which form the intermediate data of the individual mapper. Following are the steps in the mapper phase:

- **Pre-processing**: The major step is the pre-processing that processes the signals in such a way as to remove the background non-voice region from the signal to enable the effective feature extraction.

- **Feature extraction**: The second step in the mapper phase is feature extraction and these features classify the audio.

### Spectral kurtosis of the audio signal

The spectral kurtosis (SK) [40] plays a dominant role in filtering in recurring the randomly occurring signals, which are corrupted with the additive noise, and it is reported as the fourth-order cumulant of the Fourier transform (FT). Let the discrete random signal be represented as \( y(b) \), \( Q(v) \) is the \( P \)-point discrete FT and the SK at the individual frequency bin \( v \) is given as:

\[
F_1 = \frac{f_4\{Q(v), Q*(v), Q(v), Q*(v)\}}{f_2\{Q(v), Q*(v)\}^2} \tag{3}
\]

where \( Q^*(v) \) refers to the complex conjugate of \( Q(v) \), and \( f_n \) specifies the \( n^{th} \) order cumulant. The dimension of the SK is represented \([1 \times 1]\).

### Multiple kernel Mel frequency cepstral coefficient feature

The need for the multiple kernel Mel frequency cepstral coefficient (MKMFCC) feature [41] is interpreted that extracts all possible phonetic components from the audio signal through considering the low and high energy frames of the audio signal. Let us represent the MKMFCC features as \( F_2 \) and the dimension of the features is denoted as \([1 \times 13]\).

### Spectral flux of the audio signal

It is the measure of the change in the local spectrum. Let us represent the spectral flux [42] of the spectrum as \( F_3 \) and the dimension is \([1 \times 1]\).

### Delta-amplitude modulation spectrogram features

The meagre variations in the frequency and time domains of the signal are taken into account to represent the delta-amplitude modulation spectrogram (AMS) features [43]. Thus, the delta-AMS features are represented as:

\[
F_4 = F(\tau, \$) = R(\tau, \$), \Delta R_p(\tau, \$), \Delta \zeta(\tau, \$) \tag{4}
\]

where \( p \) refers to the total segments or frames, \( R(\tau, \$) \) is the AMS feature vector, and \( \Delta R_p(\tau, \$) \) is the delta-AMS feature, which is computed as:

\[
\Delta R_p(\tau, \$) = R(\tau, \$) - R(\tau - 1, \$) ; \tau = 2, \ldots, p \tag{5}
\]

The dimension of the delta-AMS features is given as \([1 \times 64]\). Therefore, the feature vector corresponding to the audio signal \( y \) is represented as:

\[
F = \{F_1, F_2, F_3, F_4\} \tag{6}
\]

where \( F_1, F_2, F_3, F_4 \) represent the SK, MKMFCC, spectral flux, and delta-AMS features and the dimension of the feature vector corresponding to the individual audio signal is \([1 \times 110]\). The output from the mapper phase is the intermediate data \( I \), and the output of the individual mappers is represented as:

\[
I = \{I_1, I_2, \ldots, I_q, \ldots, I_n\} \tag{7}
\]

where \( I_q \) corresponds to the intermediate data of the \( g^{th} \) mapper. The intermediate data is the input to the reducers in the reducer phase, which performs the audio classification.

### 3.1.2 Reducer phase

The number of the reducers equal to the total classes and the reducers are denoted as:

\[
r = \{r_1, r_2, \ldots, r_x, \ldots, r_q\} \tag{8}
\]

where \( r_x \) specifies the \( x^{th} \) reducer and there are a total of \( q \) reducers, and the class outputs from the reducers are denoted as:

\[
C = \{C_1, \ldots, C_x, \ldots, C_q\} \tag{9}
\]

where \( C_x \) is the class value of the \( x^{th} \) reducer. For the detection of the class label, the trained SVM is employed and the progressing steps of the classifier are discussed below.

**Trained SVM classifier for audio classification**

SVM [43, 45] handles the unconstrained optimization issue with the large-margin classifier aiming at the maximization of the margin in determining the optimal hyperplane. The training algorithm, G-ROA adaptively tunes the capacity,
enabling effective classification accuracy. Let the input vector $I$ with $n$ intermediate data belongs to either of the two classes, Class 1 or Class 2, represented as given in Equation (10),

\begin{equation}
(I_1, C_1), (I_2, C_2), \ldots, (I_n, C_n)
\end{equation}

where $C_x$ is the $x^{th}$ class label. The parameters of the decision variable $X(I)$ are determined at the time of learning and later during testing; the unknown patterns are classified using the following formula:

\begin{equation}
\begin{align*}
I \in \text{Class 1} & : \text{if } X(I) > 0 \\
I \in \text{Class 2} & : \text{otherwise}
\end{align*}
\end{equation}

$X(I)$ is computed based on the predefined functions of $I$.

\begin{equation}
X(I) = \sum_{x=1}^{n} w_x K(I_x, I) + w_2
\end{equation}

where $w_x$ and $w_2$ are the adjustable parameters of the classifier. The adjustable parameters $w_x$ and $w_2$ are the weight and bias of the standard SVM, which is determined based on the optimization algorithm.

$I_x$ is the training pattern and $K(I_x, I)$ refers to the kernel function. In general, the kernel function $K(I_x, I)$ exploits the exponential kernel, whereas in SVM, a new kernel function is used through summing the three kernels, polynomial, Gaussian, and RBF kernels. The advantage of using the new kernel is that the classification accuracy and the effectiveness of classification are better than the already existing kernel. The kernel is given as:

\begin{equation}
K(I_x, I) = \gamma_1 K_1 + \gamma_2 K_2 + \gamma_3 K_3
\end{equation}

where $\gamma_1$, $\gamma_2$, and $\gamma_3$ are the kernel weights that are determined using the proposed G-ROA. The polynomial kernel is represented as:

\begin{equation}
K_1 = (I_x, I + 1)^d
\end{equation}

where $d$ refers to the degree of the polynomial. The Gaussian kernel is formulated as:

\begin{equation}
K_2 = \exp \left( -\frac{\|I_x - I\|^2}{2 \sigma^2} \right)
\end{equation}

Thus, the dimensionality issue in the classifier is tackled using the Lagrangian model for which the norm needs to be minimized, maximizing the margin. The Lagrangian model is given by:

\begin{equation}
N(z, w, \gamma) = \frac{1}{2} \|z\|^2 - \sum_{x=1}^{n} [w_x [C_x X(I_k - 1)]] \text{ subject to } w_x > 0
\end{equation}

where $w_x$ specifies the Lagrange multiplier, which satisfies the condition $w_x [C_x X(I_k - 1)] = 0$. Thus, the problem is resolved by determining the minimal saddle point of $N(z, w_2, w_x)$ in terms of $z$.

**Proposed G-ROA for optimal tuning of the SVM classifier**

The ultimate goal of the proposed G-ROA is to derive the optimal kernel weights for tuning the SVM classifier, and the G-ROA follows the steps of GOA [19] with the modified equation of GOA using ROA [18]. The steps of the proposed G-ROA are demonstrated below:

**Solution encoding:** The purpose of encoding is to represent the solution vector derived using G-ROA, which is of dimension $[1 \times 3]$. The optimal values of the kernel weights, $\gamma_1$, $\gamma_2$, and $\gamma_3$ are derived using G-ROA.

**Fitness measure:** The fitness of the solutions is evaluated based on the mean square error, which is notated as:

\begin{equation}
FF = \frac{1}{n} \left[ \sum_{x=1}^{n} O_x - O_{est} \right]^2
\end{equation}

where $O_x$ is the output corresponding to the $x^{th}$ training sample and $O_{est}$ refers to the estimated output of the classifier.

**Algorithmic steps of G-ROA:** The steps of the proposed G-ROA are deliberated below:

i) **Swarm Initialization:** The population or the solution is initialized as $K_i^0; (1 \leq i \leq m)$ indicating there are $m$ grasshoppers in the search space and the dimension is referred as, $\hat{d} = (i, j)$. Additionally, the maximum $a_{max}$ and minimum $a_{min}$ values of the coefficient $a$ along with the total iterations $t_{max}$ are initialized.
ii) Evaluate the fitness of the solutions: The fitness of the solutions are evaluated based on the minimal value of the error. The fitness is determined based on Equation (19).

iii) Update the value of the coefficients: The coefficient $a$ is the effective parameter that balances the exploration and exploitation phases in such a way that $a$ decreases with the increasing iterations. The value of $a$ is computed as:

$$a = a_{\text{max}} - \tau \frac{a_{\text{max}} - a_{\text{min}}}{\tau_{\text{max}}} \quad (20)$$

iv) Update the position of the current search agent: The position of $P_{i}^{\text{th}}$ grasshopper is updated based on three factors, social interaction, gravity force, and advection of wind as in equation [33]. The standard equation of GOA is given as,

$$K_{l, \tau+1}^{\text{t}} = a \times \left( \sum_{k=1, k \neq i}^{m} a \times \frac{U_{k} - L_{k}}{2} \times s \mid K_{k}^{\text{t}} - K_{l}^{\text{t}} \mid + \frac{K_{\text{k}} - K_{l}^{\text{t}}}{D_{\text{k}}} \right) + G_{\text{t}}^{\text{t}} \quad (21)$$

where $a$ is the decreasing coefficient, $U_{\text{k}}$ and $L_{\text{k}}$ are the upper and lower bounds in $d$-dimension. The target is denoted as $G_{\text{t}}$, $D_{\text{k}}$ is the distance between the grasshoppers $l$ and $\text{k}$, $K_{k}^{\text{t}}$ specifies the position of $k_{\text{th}}$ grasshopper in $d$-th dimension. Likewise, $K_{l}^{\text{t}}$ is the position of $l_{\text{th}}$ grasshopper in $d$-th dimension. Let us assume that the dimension $d = (i, j)$ and fix $m = k = 1$, therefore, the equation becomes

$$K_{l, \tau+1}^{i,j} = a \times \left( a \times \frac{U_{i,j} - L_{i,j}}{2} \times s \mid K_{k}^{i,j} - K_{l}^{i,j} \mid \right. \left. + \frac{K_{i,j}^{i,j} - K_{l}^{i,j}}{D_{l}^{i,j}} \right) + G_{i,j}^{i,j} \quad (22)$$

The above equation reveals that the next position of the $P_{i}^{\text{th}}$ grasshopper is based on the position of the $P_{i}^{\text{th}}$ grasshopper at $\tau$, target location, and other grasshopper positions. The position update based on other grasshoppers in search space locates the search agent around the target, which is due to $a$. Additionally, the repulsion, attraction, and the comfort zones of the grasshoppers are decreased using the second $a$ in Equation (22), whereas the $a$ outside in Equation (22), balances the phases. However, the ability to deal with the multi-objectives is enabled through the integration of ROA in the above equation. The attractive and repulsive phases increase global convergence and local optimal avoidance. The bypass update rule is interpreted to modify the Equation (22) and the update rule is:

$$K_{l, \tau+1}^{i,j} = \lambda \left[ K_{l, \tau}^{i,j} \ast a(j) + K_{l, \tau}^{i,j} \ast (1 - a(j)) \right] \quad (23)$$

$$K_{l, \tau+1}^{i,j} = \lambda K_{l, \tau}^{i,j} \ast a(j) + \lambda \times K_{l, \tau}^{i,j} \ast \lambda \times K_{l, \tau}^{i,j} \ast a(j) \quad (24)$$

Assume that $\eta = i$, therefore the Equation (24) is rewritten as:

$$K_{l, \tau+1}^{i,j} = \frac{1}{\lambda} \ast a(j) \left[ K_{l+1}^{i,j} - \lambda \times K_{l+1}^{i,j} \ast \lambda \times K_{l+1}^{i,j} \ast a(j) \right] \quad (25)$$

Substitute the Equation (26) in Equation (22) as:

$$K_{l, \tau+1}^{i,j} = a \times \left( a \times \frac{U_{i,j} - L_{i,j}}{2} \times s \mid K_{k}^{i,j} - K_{l}^{i,j} \mid \right. \left. \times \left\{ \begin{array}{l}
K_{i,j}^{i,j} - 1/D_{l}^{i,j} \\
\lambda \times K_{l, \tau}^{i,j} \ast (1 + \lambda \ast a(j))
\end{array} \right\} \right) + G_{i,j}^{i,j} \quad (27)$$

The Equation (27) is the update equation of G-ROA that enhances the global optimal convergence and keeps away from converging to the local optimal solution.

v) Terminate: The steps are repeated for the maximal iterations until the global best solution is derived, and this solution represents the kernel weights of SVM.

Algorithm 1 shows the pseudocode of G-ROA.

Pseudocode of G-ROA

1. **Input**: Population $K_{i}^{\text{t}}$; $(1 \leq i \leq m)$
2. **Output**: Best position $G_{i,j}$
3. Population initialization
4. Evaluate the fitness
5. Determine $G_{i,j}$-best solution
6. While $i < \tau_{\text{max}}$
4 | RESULTS AND DISCUSSION

The results of the method are revealed in this section with an effective interpretation of the comparative results. The effectiveness is revealed through the comparison, and the significance of the method is presented as the comparative table, which displays the effective method.

4.1 | Experimental setup

The experimentation of the proposed G-ROA algorithm is performed in the MATLAB tool installed in the PC with the Windows 10 OS. Table 1 describes the simulation setup of the proposed system.

4.1.1 | Dataset 1: TUT sound event 2017 dataset

The dataset 1 [36] is accessed from TUT sound event, and it consists of the records of the acoustic scenes from the street carrying the different traffic levels along with other activities. Audio in the dataset is a subset of TUT Acoustic scenes 2017 dataset. The scene was selected as representing an environment of interest for detection of sound events related to human activities and hazard situations. A total of 24 audio records have been obtained from a single acoustic scene.

4.1.2 | Dataset 2: ESC dataset

The ESC dataset [46] is a collection of short environmental recordings available in a unified format such as 5-s long clips, 44.1 kHz, and single channel. All clips have been extracted from public field recordings available through the free sound. The image_2 is obtained from the ESC-50 dataset that is employed for the classification of the environmental sound, and this dataset comprises the labelled data with the recordings of about 2000 environmental sounds. ESC dataset consists of three parts namely ESC-50, ESC-10, and ESC-US.

ESC-50: a labelled set of 2000 environmental recordings (50 classes, 40 clips per class),
ESC-10: a labelled set of 400 environmental recordings (10 classes, 40 clips per class) (this is a subset of ESC-50 – created initially as a proof-of-concept/standardized selection of easy recordings).
ESC-US: an unlabelled dataset of 250,000 environmental recordings (5-s long clips), suitable for unsupervised pre-training.

4.2 | Experimental analysis

This section deliberates the sample input audio signals used for the audio classification. Figure 2 shows the sample results of the proposed G-ROA-SVM classifier for audio classification. Figure 2(a) and (b) shows the input audio signals 1 and 2, respectively and Figure 2(c) and (d) demonstrates the feature output using input audio signals 1 and 2 from the delta-AMS feature descriptor.

4.3 | Performance metrics

The section deliberates the metrics employed for the comparison, which is essential to deliberate the effectiveness of the effective audio classification method.

4.3.1 | Accuracy

\[
\text{Accuracy} = \frac{n^+ + p^+}{p^+ + n^+ + p^- + n^-} \tag{28}
\]

where \( p^+ \) and \( n^+ \) are the true positive and true negative and \( p^- \) and \( n^- \) are the false positive and false negative.

4.3.2 | False rejection rate

False rejection rate (FRR) is given as:

\[
\text{FRR} = \frac{n^-}{p^+ + n^-} \tag{29}
\]

4.3.3 | False alarm rate

False alarm rate (FAR) is formulated as:

\[
\text{FAR} = \frac{p^-}{p^- + n^+} \tag{30}
\]

4.4 | Competing methods

The methods used for the comparison include MultiSVM [45], CNN [16], deep belief neural network (DBN) [27], and unsupervised VAD [32]. The proposed G-ROA-SVM is
compared with the existing methods to prove the effectiveness of the proposed method.

### 4.5 | Comparative analysis

This section demonstrates the comparative analysis of the methods using two input signals taken from two datasets, and the analysis is progressed based on the training percentage and k-fold validation. The training percentage is fixed between $k = 0.4$ and 0.9, while the k-fold is set to vary between $kf = 5$ and $kf = 10$, respectively.

#### 4.5.1 | Comparative analysis using the signal_1

This section performs the analysis using the signal_1 based on the performance metrics, such as accuracy, FRR, and FAR.
Figure 3 shows the analysis using the signal_1 based on the training percentage. Figure 3(a) represents, when the training percentage, $k = 0.4$, the accuracy of the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM is $0.8309$, $0.7633$, $0.6957$, $0.7633$, and $0.8849$, respectively. Figure 3(b) represents, when the training percentage, $k = 0.4$, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FAR is $0.3203$, $0.4555$, $0.368$, $0.4555$, and $0.2697$, respectively. Figure 3(c) represents, when the training percentage, $k = 0.4$, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FRR of $0.0126$, $0.0137$, $0.0146$, $0.0137$, and $0.0292$, respectively.

Figure 4 shows the analysis using the signal_1 based on the $k$-fold. Figure 4(a) shows, when the $k$-fold, $k_f = 5$, the accuracy of the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM is $0.5386$, $0.5928$, $0.7013$, $0.888$, and $0.948$, respectively. Figure 4(b) shows, when the $k$-fold, $k_f = 5$, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FAR of $0.4613$, $0.4071$, $0.2986$, $0.112$, and $0.052$, respectively. Figure 4(c) shows, when the $k$-fold, $k_f = 5$, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FRR of $0.5613$, $0.5071$, $0.3986$, $0.212$, and $0.152$, respectively.

4.5.2 | Comparative analysis using the signal_2

Figure 5 shows the analysis using the signal_2 based on the training percentage. Figure 5(a) shows, when the training percentage, $k = 0.4$, the accuracy of the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM is $0.8166$, $0.7899$, $0.8509$, $0.7962$, and $0.9083$, respectively. Figure 5(b) shows, when the training percentage, $k = 0.4$, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FAR is $0.1746$, $0.21$, $0.1491$, $0.2072$, and $0.0873$, respectively. Figure 5(c) shows, when the training percentage, $k = 0.4$, the methods MultiSVM, CNN, DBN,
unsupervised VAD, and G-ROA-SVM acquired the FRR is 0.023, 0.31, 0.2491, 0.2417, and 0.0115, respectively.

Figure 6 shows the analysis using the signal_2 based on the k-fold. Figure 6(a) shows, when the k-fold, kf = 5, the accuracy of the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM is 0.7597, 0.928, 0.685, 0.888, and 0.954, respectively. Figure 6(b) shows, when the k-fold, kf = 5, the methods MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FAR as 0.2402, 0.072, 0.3149, 0.112, and 0.046, respectively. Figure 6(c) shows, when the k-fold, kf = 5, the methods, MultiSVM, CNN, DBN, unsupervised VAD, and G-ROA-SVM acquired the FRR is 0.3402, 0.172, 0.4149, 0.212, and 0.146, respectively.

4.6 | Comparative discussion

The comparative discussion of the methods is revealed in Table 2, which reveals the effectiveness through the performance metrics. The performance is best at the training percentage and k-fold of signal_1 at k = 0.9 and kf = 10.

The processing speed of the proposed system G-ROA-SVM is compared with other existing methods given in Table 3. The time and speed-up ratio of the proposed system with other parallel computing methods are given in Table 4. The speed of the proposed system with the map-reduce platform is higher than all other parallel computing methods.

4.7 | Statistical analysis

Table 5 illustrates the statistical analysis based on training percentage and k-fold with accuracy, FAR, and FRR methods, and the mean and variance of MultiSVM, CNN, DBN, unsupervised VAD, and proposed G-ROA-SVM are determined. Based on the training percentage, the accuracy of the proposed G-ROA-SVM in case of mean is 0.9580, whereas the accuracy of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.8983 0.8820, 0.8575, and 0.8818, respectively. The
FIGURE 5 Comparative analysis based on training percentage: (a) accuracy, (b) FAR, and (c) FRR. FAR, false alarm rate; FRR, false rejection rate

Accuracy of the proposed G-ROA-SVM in terms of variance is 0.0002, whereas the accuracy of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0002, 0.0003, 0.0002, and 0.0005, respectively. The FAR of the proposed G-ROA-SVM in case of mean is 0.0680, whereas the FAR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.1364, 0.1668, 0.1013, and 0.0949, respectively. The FAR of the proposed G-ROA-SVM in terms of variance is 0.0004, whereas the FAR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0004, 0.0005, 0.0003, and 0.0002, respectively. The FRR of the proposed G-ROA-SVM in case of mean is 0.0117, whereas the FARR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0004, 0.0049, 0.0006, and 0.0046, respectively. The FRR of the proposed G-ROA-SVM in terms of variance is 0.0002, whereas the FRR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0003, 0.0001, 0.0001, and 0.0004, respectively. Based on the k-fold, the accuracy of the proposed G-ROA-SVM in case of mean is 0.9550, whereas the accuracy of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.7500, 0.9315, 0.7445, and 0.9560, respectively. The accuracy of the proposed G-ROA-SVM in terms of variance is 0.0002, whereas the accuracy of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0003, 0.0005, 0.0001, and 0.0004, respectively. The FAR of the proposed G-ROA-SVM in case of mean is 0.0217, whereas the FAR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0419, 0.0618, 0.3060, and 0.9560, respectively. The FRR of the proposed G-ROA-SVM in terms of variance is 0.0003, whereas the FRR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0001, 0.0002, 0.0004, and 0.0002, respectively. The FRR of the proposed G-ROA-SVM in case of mean is 0.0023, whereas the FRR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.3493, 0.1677, 0.3552, and 0.0026, respectively. The FRR of the proposed G-ROA-SVM in terms of variance is 0.0004, whereas the FRR of the existing MultiSVM, CNN, DBN, and unsupervised VAD is 0.0004, 0.0003, 0.0002, and 0.0001, respectively.
FIGURE 6 Comparative analysis based on k-fold: (a) accuracy, (b) FAR, and (c) FRR. FAR, false alarm rate; FRR, false rejection rate.

TABLE 2 Comparative discussion of signal_1 training percentage and k-fold at $k = 0.9$ and $k_f = 10$

| Based on training percentage | Methods | MultiSVM | CNN | DBN | Unsupervised VAD | G-ROA-SVM |
|-------------------------------|---------|----------|-----|-----|------------------|-----------|
| Accuracy                      | 0.8985  | 0.8823   | 0.8577 | 0.8823 | 0.9600            |
| FAR                           | 0.1368  | 0.1673   | 0.1016 | 0.0951 | 0.0684            |
| FRR                           | 0.0007  | 0.0050   | 0.0007 | 0.0050 | 0.0119            |

| Based on k-fold               | Methods | MultiSVM | CNN | DBN | Unsupervised VAD | G-ROA-SVM |
|-------------------------------|---------|----------|-----|-----|------------------|-----------|
| Accuracy                      | 0.7503  | 0.9320   | 0.7446 | 0.9600 | 0.9600            |
| FAR                           | 0.0420  | 0.0620   | 0.3064 | 0.1020 | 0.0220            |
| FRR                           | 0.3497  | 0.1680   | 0.3554 | 0.0027 | 0.0027            |

Abbreviations: CNN, convolutional neural network; DBN, deep belief neural network; FAR, false alarm rate; FRR, false rejection rate; G-ROA-SVM, grasshopper-ride optimization algorithm-support vector machine; SVM, support vector machine; VAD, voice activity detection.
and is complex to be handled, and hence, the big data is handled perfectly using MRF, which comprises the feature extraction and audio classification steps. The features are progressed using the proposed G-ROA-based SVM. The optimal tuning of the SVM enhances the effectiveness of classification, and the analysis of the effectiveness is performed using the data taken from the TUT sound event 2017 dataset and ESC dataset. The environmental sounds are easily recognized using the developed classifier and the robustness is enhanced. The analysis of the proposed method in comparison with the existing methods reveals that the proposed method outperformed with the maximal accuracy of 0.96, minimal FAR and FRR of 0.022 and 0.0119, respectively. The future dimension of the audio classification relied in using any deep learning networks for audio classification.

## 5 | CONCLUSIONS

The audio classification using the SVM classifier enables the accurate classification and enhances the robustness towards the noise. The audio database exhibits the big data configuration and is complex to be handled, and hence, the big data is handled perfectly using MRF, which comprises the feature extraction and audio classification steps. The features are progressed using the proposed G-ROA-based SVM. The optimal tuning of the SVM enhances the effectiveness of classification, and the analysis of the effectiveness is performed using the data taken from the TUT sound event 2017 dataset and ESC dataset. The environmental sounds are easily recognized using the developed classifier and the robustness is enhanced. The analysis of the proposed method in comparison with the existing methods reveals that the proposed method outperformed with the maximal accuracy of 0.96, minimal FAR and FRR of 0.022 and 0.0119, respectively. The future dimension of the audio classification relied in using any deep learning networks for audio classification.

### REFERENCES

1. Wu, Y., Mao, H, Yi, Z.: Audio classification using attention-augmented convolutional neural network. Knowl. Based Syst. 161, 90–100 (2018)
2. Akagi, M., et al.: Toward affective speech-to-speech translation: strategy for emotional speech recognition and synthesis in multiple languages. In: Proceedings of Signal and Information Processing Association Summit and Conference, 1–10. Siem Reap, Cambodia (2015)
3. Hansen, J.H, Liu, G.: Unsupervised accent classification for deep data fusion of accent and language information. Speech Commun. 78, 19–33 (2016)
4. Cai, R., et al.: Highlight sound effects detection in audio stream. In: Proceedings of IEEE International Conference on Multimedia and Expo, Baltimore, MD, USA (2003)
5. Baele, M., Biermacki, C., Greff, R.: Real-time monophonic and polyphonic audio classification from power spectra. Pattern Recognit. 92, 82–92 (2019)
6. Clavel, C., Ehrette, T, Richard, G.: Events detection for an audio-based surveillance system. In: Proceedings of IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands (2015)
7. Zubair, S., Yan, F, Wang, W.: Dictionary learning based sparse coefficients for audio classification with max and average pooling. Digit. Signal Process. 23(3), 960–970 (2013)
8. Duxans, H., Anguera, X., Conejero, D.: Audio-based soccer game summarization. In: 2009 IEEE International Conference on Broadband Multimedia Systems and Broadcasting (BMSB), Bilbao, Spain (2009)
9. Zlatintsi, A., et al.: Audio salient event detection and summarization using audio and text modalities. In: Proceedings on 23rd European Signal Processing Conference (EUSIPCO), Nice, France (2015)
10. Li, X., et al.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. IEEE Trans. Geosci. Remote Sens. 52(11), 7086–7098 (2014)
11. Karthikeyan, N., Sangeetha, S., Krishnamoorthy, N.: Performance evaluation of optimization algorithm using scheduling concept in grid environment. IIOAB J. 7(9), 315–323 (2016)
12. Vidyadhar, C., Sandhya, N., Premchand, P.: A semantic word processing using enhanced cat swarm optimization algorithm for automatic text clustering. Multimed. Res. 2(4), 23–32 (2019)
13. Zubair, S., Wang, W.: Audio classification based on sparse coefficients. In: Sensor Signal Processing for Defence, London (2011)
14. He, L., et al.: Time-frequency feature extraction from spectrograms and wavelet packets with application to automatic stress and emotion classification in speech. In: Proceedings of International Conference on Information Communications and Signal Processing, pp. 1–5 (2019)
15. JadHAV, P.P., Joshi, S.D.: Fractional weightage based objective function to a hybrid optimization algorithm for model transformation. Evolut. Intellig. 1–13 (2018)
16. Phan, H., et al.: Improved audio scene classification based on label-tree embeddings and convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25(6) (2017)
17. Sawhney, N., Macs, P.: Situational awareness from environmental sounds. Technical Report. Massachusetts Institute of Technology (1997)
18. Clarkson, B., Sawhney, N., Pentland, A.: Auditory context awareness via wearable computing. Energy. 400(600), 20 (1998)
19. Walekar, S., Saha, G.: Classification of audio scenes with novel features in a fused system framework. Digit. Signal Process. 75, 71–82 (2018)
20. Zhang, M., et al.: Sparse coding for sound event classification. In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 1–5. Kaohsiung, Taiwan (2013)
21. Wang, J.C., et al.: Robust environmental sound recognition for home automation. IEEE Trans. Autom. Sci. Eng. 5(1), 25–31 (2008)
22. Hinton, G.E., et al.: Supervised model training for overlapping sound events based on unsupervised source separation. In: Proceedings of 2013 IEEE International Conference on Acoustics, Speech And Signal Processing, 8677–8681, BC, Canada (2013)
23. Srinivas, V., Santhirani, Ch.: Hybrid particle swarm optimization-deep neural network model for speaker recognition. Multimedia Res. 3(1), 1–10 (2020)
24. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio based event detection for multimedia surveillance. IEEE Int. Conf. Acoust. Speech Signal Process. Proc. 5 (2006)
25. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29(6), 82–97 (2012)
26. Ghabahi, O., Hernando, J.: Deep belief networks for i-vector based speaker recognition. In: Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (1) 1700–1704. Florence, Italy (2014)
27. Wang, C., et al.: Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network. IEEE/ACM Trans. Audio Speech Lang. Process. 26(8), 1336–1351 (2018)
28. Fan, W., Han, Z., Wang, R.: An evaluation model and benchmark for parallel computing frameworks. Mob. Inf. Syst. 7, 1–14 (2018)
29. Nanni, L., et al.: Combining visual and acoustic features for audio classification tasks. Pattern Recognit. Lett. 881, 49–56 (2017)
30. Binu, D., Karivayapp, B.S.: Ride, N.N.: A new rider optimization algorithm-based neural network for fault diagnosis in analog circuits. IEEE Trans. Instrum. Meas. 99, 1–25 (2018)
31. Saremi, S., Mirjalili, S., Lewis, A.: Grashopper optimisation algorithm: theory and application. Adv. Eng. Softw. 105, 30–47 (2017)
32. Ali, Z., Talha, M.: Innovative method for unsupervised voice activity detection and classification of audio segments. IEEE Access, 6, 15494–15504 (2018)
33. Souli, S., Lachiri, Z.: Audio sounds classification using scattering features and support vectors machines for medical surveillance. Appl. Acoust. 130, 270–282 (2018)
34. Hong, K.J., et al.: A robust audio classification system for detecting pulmonary edema. Biomed. Signal Process. Control. 46, 94–103 (2018)
35. Arumugam, M., Kalapanna, M.: Feature selection based on MBFOA for audio signal classification under consideration of Gaussian white noise. IET Signal Process. 12(6), 777–785 (2018)
36. TUT Sound Events 2017, Evaluation Dataset. https://zenodo.org/record/1040179?persistentId=doi:10.4441/2018/047. Assessed 30 Aug 2018
37. Singh, S.P., Jaiswal, U.C.: Machine learning for big data: a new perspective. Int. J. Appl. Eng. Res. 15(3), 2753–2762 (2018)
38. Leng, Y., et al.: Audio scene recognition based on audio events and topic model. Knowl. Based Syst. 125, 1–12 (2017)
39. Zubair, S., Yan, F., Wang, W.: Dictionary learning based sparse coefficients for audio classification with max and average pooling. Digit. Signal Process. 23, 960–970 (2013)
40. Antoni, J.: The spectral kurtosis of nonstationary signals: formalisation, some properties, and application. In: Proceedings of the 12th European Signal Processing Conference, 6–10. Vienna, Austria (2004)
41. Ramiah, V.S, Rao, R.R.: Speaker diarization system using MKMFCC parameterization and WLI-fuzzy clustering. Int. J. Speech Technol. 19(4), 945–963 (2016)
42. Su, L., Yang, Y.: Power-scaled spectral flux and peak-valley group-delay methods for robust musical onset detection. In: Proceedings of International Computer Music Association (2014)
43. Haridas, A.V., Marimuthu, R., Chakraborthy, B.: A novel approach to improve the speech intelligibility using fractional delta-amplitude modulation spectrogram. Cybern. Syst. 49(7), 421–521 (2018)
44. Wang, J., et al.: Content-based audio classification using support vector machines and independent component analysis. In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR’06), Hong Kong, China (2006)
45. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT), 92, 144–152. Pittsburgh, PA, USA (1992)
46. ESC Dataset for Environmental Sound Classification. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YDEPUT. Assessed 30 Aug 2018

How to cite this article: Singh, S.P., Jaiswal, U.C.: Audio classification using grasshopper-ride optimization algorithm-based support vector machine. IET Circuits Devices Syst. 15(5), 434–447 (2021). https://doi.org/10.1049/cds.12039