IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2022, accepted 22 July 2022, date of publication 16 August 2022, date of current version 25 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3198966

RESEARCH ARTICLE

Gunshots Localization and Classification Model Based on Wind Noise Sensitivity Analysis Using Extreme Learning Machine

SHAHZAD AHMAD QURESHI[1], LAL HUSSAIN[2,3], HAYA MESFER ALSHAHRANI[4], SYED RAHAT ABBAS[1], MOHAMED K NOUR[5], NAYABB FATIMA[1], MUHAMMAD IMRAN KHALID[1], HUNIYA SOHAIL[1], ABDULLAH MOHAMED[6], AND ANWER MUSTAFA HILAL[7]

[1]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad 45650, Pakistan
[2]Department of Computer Science and Information Technology, King Abdullah Campus Chatter Kalas, University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir 13100, Pakistan
[3]Department of Computer Science and Information Technology, Neelum Campus, University of Azad Jammu and Kashmir, Athmuqam, Azad Kashmir 13230, Pakistan
[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[5]Department of Computer Sciences, College of Computing and Information System, Umm Al-Qura University, Mecca 24382, Saudi Arabia
[6]Research Centre, Future University in Egypt, New Cairo 11745, Egypt
[7]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Al-Kharj 16278, Saudi Arabia

Corresponding authors: Shahzad Ahmad Qureshi (drsaqureshi@pieas.edu.pk), Lal Hussain (lal.hussain@ajku.edu.pk), and Anwer Mustafa Hilal (a.hilal@psau.edu.sa)

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R237), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4310373DSR20). This work was supported in part by the Higher Education Commission (HEC), Pakistan; in part by Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan, under Grant 03-6-1-004-2016.

ABSTRACT The gunshot event localization and classification have numerous real-time applications. The study is also useful for steering the video camera and guns in the directed direction. This paper proposes a framework that can be used for a surveillance system to accurately localize and classify the type of gunshots impregnated with wind noise. The main contribution of this paper is the localization of the gunshot for the very first time using Hadamard product with wavelet de-noising in windy conditions. We have evaluated our framework on airborne gunshots acoustic dataset, and a derived (simulated) sound dataset, as an offline scenario, using four microphones' geometry. For localization, the proposed system outperformed with an accuracy of 99.95%. The other contribution is a sensitivity-based comprehensive examination of gunshot sound signals, with normal to strong wind noise of varying SNRs, for machine learning and deep learning classifiers to categorize the type of gunshots. For classification, it has been found, not known before for the gunshots dataset, that ELM is robust for original, normal, and strong windy environments with an accuracy of 93.01%, 91.61%, and 88.11% respectively with the threshold SNR. A comprehensive comparison of recent techniques with the proposed approach has also been added.

INDEX TERMS Acoustic signal processing, azimuth, direction of arrival, elevation angle, extreme leaning machine, phase delay, time delay.

I. INTRODUCTION

Within the past few decades, a renowned research area is the localization and categorization of acoustic events in numerous fields like fine particle detection [1], medical health sciences [2], and audio surveillance systems [3] including challenging actual time environments [4]. The array of microphones is used for acquiring valuable signals with time-to-live (TTL) variations [5]. The study of gunshot events localization and classification has important implications in real-time

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

surveillance systems. It has gained interest for steering the video cameras, and even guns, in the directed directions. Similarly, the classification of gunshot events is important along with its localization. Therefore, the positional parameters and related classification of gunshots motivate to view the scene using video devices in a surveillance system, especially in the case of wind noise-contaminated environments.

The acoustic emission is its source positioning or localization and classification (L&C) linked to complex nonstationary waves with detailed source information. Initially, the acoustic signals are detected followed by time delay (TD) localization of the source. The localization is carried out in real-time by finding the cross-correlation between acoustic signals acquired for each microphone based on a specific common (reference) one. The azimuth and elevation angles as direction of arrival (DOA) metrics are estimated using TD measurement techniques. Some researchers working on acoustic signals used the frequency domain in finding the DOA. Astapov *et al.* [6] selected the urban and military security system for acoustic event localization based on gunshots. They worked on a circular array of unmanned group sensors (UGS) for the estimation of the direction of arrival of gunshot acoustic events with a computationally cost-effective solution. Gaikwad *et al.* [7] worked on combat operations support by means of localization of enemy troops in real-time which was considered beneficial to planning a war strategy. Their enemy localization method was based on triangulation for an enemy localization, i.e. acoustic source positioning, using two microphones and a single acoustic source strategy. They also reported four different complex scenarios using different stages. Valenzise *et al.* [3] showed an acoustic surveillance system in their remarkable work that identified anomalous acoustic events and localized the event source with camera steering in the predicted direction. They used the least-squares-based localization algorithm to compute the azimuth and elevation angles to find the time difference of arrival using an array of microphones. Astapov *et al.* [8] based their research on gunshot shockwave and muzzle blast direction of arrival on civilian as well as military security systems. They used circular microphone geometry of microphones and reported an adequate direction of arrival time in their published work with computational low overhead. Pathrose *et al.* [9] worked on the localization problem to ascertain the attack direction by small firearms for retaliation. Consequently, the shooting position determination for surveillance is important to tackle any possible attacks. They used multiple microphones located at the same distance to acquire the time of arrival for different acoustic events and determine the positional information as azimuth and elevation angles of the acoustic source with respect to the corresponding topology of microphones.

Numerous ideas and approaches have been extensively employed for acoustic event classification using learning-based or data-driven methods to develop the mapping function between unknown sound signals and the predicted class label. The discriminative features are extracted for localization by involving massive amounts of sensor data with the help of artificial intelligence-based tools, like artificial neural networks (ANN), and deep neural networks (DNN). The localization was dealt with as a regression job finding distance using a model, and finally representing the position-related details by determining the missing and uncertain knowledge of the physical environment. Anzai [10] allocated machine learning (ML) models for classification tasks by introducing the structure of patterns. Efforts were made by many researchers to optimize the ML algorithms. Bottuo *et al.* [11] wrote a review for text classification highlighting optimized ML algorithms using case studies. They mentioned DNN application in feature space to find the optimal solution. Similarly, Huang *et al.* [12] used single hidden layer feedforward neural network (SLFNs) for introducing an extreme learning machine (ELM) where branch weights for the inputs are randomly selected and the outputs are estimated in an analytic manner. It has a fast learning capability as compared with ML, including its robust and consistent problem-solving ability to overcome the overfitting problem during the training phase in conventional neural networks [13]. Corriea *et al.* [14] employed deep-feed forward neural networks for the localization of acoustic events. It is based on the training of multiple neural networks under the given conditions. Vera-Diaz *et al.* [15] employed convolutional neural networks for acoustic source localization using massive positioning data to carry out the appropriate learning during the training phase. Here, it is important to note that deep learning application requires a sufficient number of instances to automatically discover the representations which are not always possible in sensitive cases [16].

Numerous cohorts solved acoustic event categorization problems using ML strategies. The Mel-frequency cepstral coefficient (MFCC) based features have been commonly adopted for acoustic events [17]. The principal component analysis (PCA) with its linear and kernelized variants has been used to reduce the features to a more discriminative form [18]. The classifiers that have been commonly used for acoustic event classification are naïve Bayes (NB), k-nearest neighbors (k-NN), support vector machine (SVM), linear discriminant analysis (LDA), and random forest (RF) in individual (traditional) as well as hybrid (or ensemble) capacity [19]. SVM is based on both structural and empirical risk minimizations. The former is based on a boundary creation in a way that maximizes the margin between the classes whereas the latter minimizes the number of misclassifications between the classes by transforming the input space into high dimensional space that leads to generalization improvement [20]. NB classifier makes feature variables independent, so all variations and characteristics of each class can be learned [21]. The *k*-NN classifier is the simplest algorithm as it does need any information about the existing data. It is a non-parametric learning algorithm and it is based on the nearest neighbors to the unknown instances with features characteristic of either of the classes [22]. The RF classifier [23] is an ensemble-based technique in which multiple decision trees are used as base learners. Improvements equivalent to more than doubling the data can be achieved using RF algorithms offering better results from the same data size. Recently, the active learning technique has been used with a modified breaking ties algorithm with multinomial logistic regression for the classification of hyperspectral images of aerial views using satellite images [16]. Liu *et al.* [24] carried out acoustic signals-based fault detection on belt-conveyor idlers. After the acoustic signal acquisition, features were extracted by MFCC and then applied to different machine learning algorithms. Dabetwar *et al.* [25] worked on ultrasonic data acquisition for multiple structural health monitoring systems, and a classical supervised machine learning algorithm was applied to determine the damage levels of different signals. Similarly, Pham *et al.* [26] detailed a survey on machine learning, deep learning, and federal learning applied in the field of intelligent radio-signal processing. Rozemberczki *et al.* [27] developed a deep learning framework with a combination of machine learning for solving the problem of spatiotemporal signals. The main focus of this research was to build temporal geometric deep- and machine-learning models in a unified form.

Shi *et al.* [28], [29] introduced Gammatone Frequency Cepstral Coefficients (GFCC) as an alternative to MFCC [30] for speaker recognition systems. They replaced Mel filter bank with the gammatone filter bank (GFB) to improve robustness. They used multitaper estimation, MVA (mean subtraction, variance normalization, and autoregressive moving average filter) to generate GFB. Similarly, Thiruvengatanadhan [31], [32] introduced Power Normalized Cepstral Coefficients (PNCC) as the acoustic features. In this feature extraction technique, the discrete wavelet transform (DWT) based features [33], [34] are grouped into k number of groups using k-means clustering The classification is based on the minimum distance between the cluster centroid and the feature vector. Some cohorts also worked on the hybrids [35], [36], [37] of these acoustic feature extraction techniques and found improved results that would have been obtained otherwise on an individual basis.

The event localization, as well as classification metrics, are usually affected due to the distortion of the sound signal when it passes through the wind medium [38]. The investigation of degradation to SNR of the sound signal was successfully carried out as a real-world problem by white noise analysis. The noise present in the acoustic signals has been analyzed by numerous filters, like Savitzky-Golay [39] and the moving average filters [40]. The performance of the conventional filters underperforms by changing the signal parameters, while the behavior of the adaptive filters is repeatedly reflected in the statistical properties of the signals including noise [41]. Bhoyar *et al.* employed recursive least square (RLS) and least mean square (LMS) filters for noise cancellation to measure the precise signal [42]. Dhimane *et al.* [43] conducted a study on adaptive filter usage for their comparison concerning stability, efficiency, and computational cost for various applications. Similarly, Khan *et al.* [44] carried out a survey to compare the RLS, notch, and LMS filters. They found that the RLS relatively performed better than the rest of the two, but suffered an additional cost. For an acoustic noise distribution, Goubran *et al.* [45] employed adaptive type filters for the suppression of noise and enhanced the vehicle sound signals. Breining *et al.* [46] used adaptive filters of complex types with high order for the acoustic echo control. The process suffered from additional computational costs. Thenua *et al.* [47] worked on lung sound signals with the application of adaptive filters on bio-signals and proposed a novice algorithm for the classification of acoustic events.

We focused on normal and strong wind noise models as the environmental noise type in this paper. Further, we have analyzed distinctive levels of windy sound, by observationally finding parametric thresholds for different filters. In common, a specific filter outflanks up to a certain limit with the acoustic events sifted utilizing conventional, as well as adaptive filters prior to L&C to check the execution of the framework under normal and extreme noise-contaminated acoustic signals.

The main contributions of this work are summarized as follows:

- A framework is proposed for the localization and classification of gunshots in windy conditions.
- Comparison of conventional filters with adaptive filters for gunshot signals with wind noise for varying SNR.
- The Hadamard product with wavelet de-noising is used for the very first time to localize gunshots in windy conditions.
- A comprehensive sensitivity analysis of gunshot signals, with normal to strong wind noise of varying SNR, for machine learning and deep learning classifiers to categorize the type of gunshots.
- Comparison of the proposed work with other known existing research works for localization and classification of acoustic sources.

The organization of this article is as follows: **Section** II illustrates the materials and methods, **Sections III** is based on results and discussion, while **Section IV** summarizes the paper with conclusions.

II. MATERIALS AND METHODS

The establishment of sound waves is based on the propagation of acoustic waves. The sound waves need mechanical vibrations to travel through the flexible medium. Sound waves move faster in solids, while the velocity of sound is relatively slow in gases and liquids congruent to their respective degree of compactness on the atomic scale [48]. The proposed system for sound source localization and classification used for the detection of acoustic events is illustrated in **Figure** 1, where **Figure** 1 (a) represents the data flow between different modules, and **Figure** 1 (b) illustrates the detailed workflow of complete acoustic localization and classification system including the impregnation with noise, filtering prior to the L&C phases for measuring and analyzing the acoustic signals. The analysis of different L&C systems is carried out after the noise addition and filtering of the acoustic signals.

A. DATASET

The airborne gunshots dataset [49] (AGD-2021) has been used in this research article which contains different categories of weapons with 722 sound effects as illustrated in **Table** I. The dataset of gunshots was recorded in an unknown acoustic environment as laid down by AGD-2021. We reconstructed sound effects recordings with existing wind noise models [49] and simulated the time delay of arrival at particular azimuth (ϕ) and elevation (θ) angles. The detail of the class imbalance of the dataset is shown in **Section** II (E) (1), whereas the visualization of the feature space of MFCC is illustrated by the t-SNE plot (**Section** III (A).

B. PREPROCESSING

For localization, the notion of preprocessing is to generate the data for four microphones' geometry before normal and strong wind noise simulation is carried out. For classification, the Airborne dataset is directly simulated for reverberant wind noise conditions. The high-class imbalance in this dataset is alleviated to improve generalization as illustrated in **Section** 2.5.1. Extracting unique features from the acoustic signals, viz. original, normal, and strong wind noise impregnated signals is explained in **Section** 2.5.2 by using MFCC vectors.

1) SIMULATION FOR LOCALIZATION

The original signals (Airborne dataset) are employed for the simulation of acoustic signals (localization dataset) using a four-node omnidirectional geometry at known, 70 cm, positions as illustrated in **Figure** 2 [50]. Node 1, viz. Microphone-1 (Mic-1) has been selected as the reference node. In practical applications, the data acquisition is nearly continuously accompanied by wind noise impregnation. For noise impregnation, original sound signals are treated with a noise model [38] for normal as well as intense wind conditions. The preprocessing of signals for the detection of noise falls apart its quality.

C. ANALYSIS OF FILTERS FOR NOISE

The elimination of noise or outliers from the sound events is analyzed using filters. The filter behavior is found corresponding to the limiting value of SNR for the simulation of strong and normal wind models using conventional filters with their adaptive counterparts before the localization or classification phases.

1) CONVENTIONAL FILTRATION

This type of filtration is characterized by filtering out the detrimental components of the signals in the neighborhood by using fixed parameters on conventional filters. The notion is to partition the acquired signal into multiple portions as illustrated in Figure 3. The sequential application of filter then follows signal reconstruction using the filtered sections.

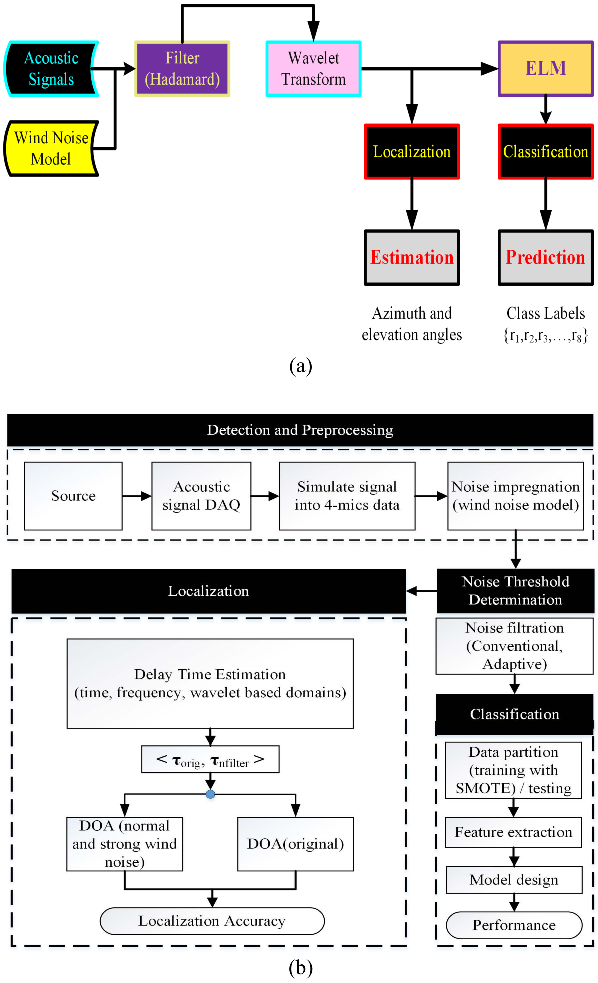

Acoustic Signals
Wind Noise Model
Filter (Hadamard)
Wavelet Transform
ELM
Localization
Classification
Estimation
Prediction
Azimuth and elevation angles
Class Labels $\{r_1, r_2, r_3, \ldots, r_8\}$
(a)
Detection and Preprocessing
Source
Acoustic signal DAQ
Simulate signal into 4-mics data
Noise impregnation (wind noise model)
Noise Threshold Determination
Localization
Delay Time Estimation (time, frequency, wavelet based domains)
$< \tau_{orig}, \tau_{nfilter} >$
DOA (normal and strong wind noise)
DOA(original)
Localization Accuracy
Noise filtration (Conventional, Adaptive)
Classification
Data partition (training with SMOTE) / testing
Feature extraction
Model design
Performance
(b)

FIGURE 1. Proposed framework for acoustic event localization and classification, (a) data flow diagram, (b) L&C framework.

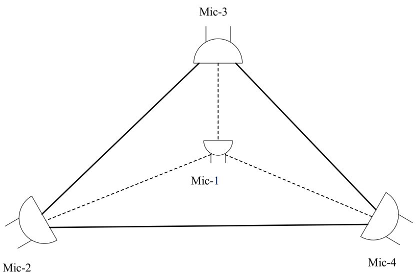


FIGURE 2. Visualization of 4-microphone geometry in an omnidirectional manner (each one of the mics is 70 cm apart from one another).

An overview of filters used in this work is given with a brief working principle explained as given by:

The **weighted average filter** is based on the application of a signal that has been segmented, taking the mean of N sequential segments of a waveform [51]. The reconstructed signal smoothness is attributed to the removal of noise, and it is given by $y[n] \overset{\text{def}}{=} \sum_{k=0}^{N-1} x[n-k]$.

In the **median filter**, a window scope is defined for ordered samples where the central value is thought to be the output.

TABLE 1. Airborne dataset (gunshots).

Class	Label	Class cardinality
Handgun, pistol and revolver	1	47
Handgun, pistol and semi-automatic	2	179
Rifle and bolt action	3	140
Rifle and carbine	4	111
Rifle and fully automatic	5	124
Rifle and lever action	6	24
Shotgun and bolt action	7	31
Shotgun and pump action	8	66

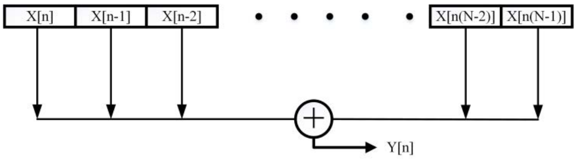


FIGURE 3. Segmentation of an acoustic signal.

The removal of outliers results in the smoothness of the signal. As a non-linear method, it is a key filter to isolate the impulsive noise while not affecting the useful components in the signal.

Savitzky–Golay filter is defined in the time domain-based moving window with least squares used for polynomial fitting [52]. The waveform obtained is smooth and represented as $g_m = \sum_{k=-n_L}^{n_R} c_{k+n_L} s_{m+k}$. Here, $\mathrm{g_m}$ represents the output signal, s_{m+k} is the input signal, the midpoint is m, n_R represents the point-count to the right of m, and n_L is the point-count to the left.

2) ADAPTIVE FILTRATION

There exist multiple unknown parameters in a real-time system with non-linearly dynamic variations and we cannot rely only on rigid traditional digital filters. Hence, adaptive filters are imperative for varying environments. The main aim of these filters is to minimize the cost function between the output of the adaptive filter and desired signal to achieve optima. Various cost functions have been proposed so far to optimize these filters. Another interesting aspect of adaptive filters is the way they merge the distributed signals. Combine then adapt and adapt then combine are commonly practiced approaches for merging signals [53]. In adaptive filters, the parameters of the input signal are derived by processing and updating them [41]. The working principle of an adaptive filter is illustrated in **Figure** 4. Here, the error between the target and filtered output is represented by e(n), x(n) are the input noise signals, d(n) are the target signals, and the output is represented by y(n).

An overview of adaptive filters used in this work is given by:

A **fair cost function-based adaptive filter** has been proposed by Guan *et al.* [53]. Spatial and temporal weights

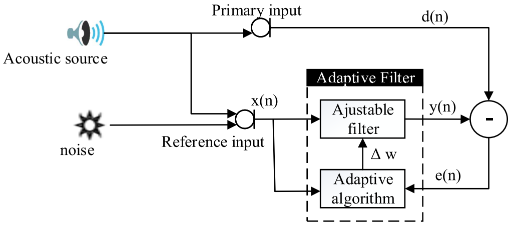


FIGURE 4. Working principle of an adaptive filter.

of varying parameters are computed for the cost function. The steepest descent method is utilized to update the weight vector. ATC technique is resorted to combining the temporal details of system parameters. Let $\boldsymbol{V}^0$ is the initial weight vector of the input signal $A\,(i)$ with variance v of varying parameters. At any instance i the loss $L(i)$ is defined as: $L\,(i) = \boldsymbol{V}^0 A\,(i) + v\,(i) - \boldsymbol{V}\,(i)\,A(i)$. The fair cost function C(i) based on $L(i)$ is computed by:

$$\mathrm{C}\,(i) = \gamma^2 \left[\frac{|L\,(i)|}{\gamma} - log(1 + \frac{|L\,(i)|}{\gamma} \right] \tag{1}$$

where γ is the threshold in such a way that $\gamma > 0$. Similarly, the adaptive Kalman filter (AKF) is presented by [54] to model the relation between gyroscope random noise and white noise, providing the fact that the gyroscope noise model is continuous while the Kalman filter is digital in the time domain. Therefore, the noise model is linked to the filter in the continuous-time domain as a first step, and then it is converted into a discrete form to connect with the digital Kalman filter. This filter performs better for both Gaussian noise as well as color noise. The adaptive filter for non-Gaussian flicker noise (NGFN) proposed by Parshin and Parshin [55] endeavored to gain balance between the energy of the signal and the width of the spectrum to get the best signal-to-noise ratio. It uses the adaptive Bayesian technique to compute the variables of NGFN.

The **least mean square (LMS) filter** uses optimized filter weighting factors, and its working principle is the stochastic gradient strategy [52] trying to behave like the desired signal. The objective is achieved by finding the error based on least mean squares by varying the filter coefficients, as illustrated in **Figure** 5. The adjustable parameters are: y(n) represents the filtered output, d(n) is the target signal, e(n) represents the error between target and filtered outputs, w is the weight, h represents filtering, and x(n) is the noisy input signal.

The **recursive least square (RLS) filter** is based on the minimization of the least mean square error and handles the coefficients of the filter in a recursive manner [47]. Further, the filter outclasses in performance at an additional computational overhead [56]. The principle of the RLS filter is illustrated in **Figure** 6. The parametric updating is carried out recursively, where e(n) is the error between the predicted and the target outputs, d(n) is the target output, d'(n) is the predicted output, and Δw is the weight- update to the signal coefficients.

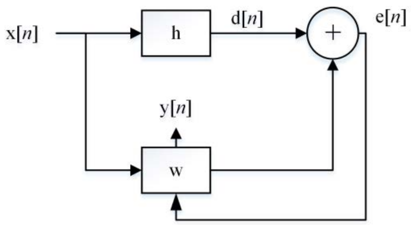


FIGURE 5. Working principle of LMS filter.

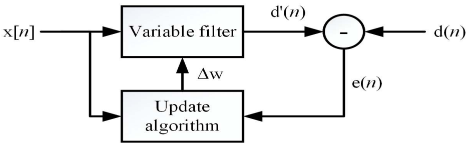


FIGURE 6. Working principle of RLS filter.

In the **amplitude-based frequency filters (ABFF)**, the extraction of frequency components that are useful, as depicted in **Figure** 7, is conducted. Here, x'(n) represents the input noisy signals, x(n) indicates the target signal, e(n) is the error existing across the target and filtered outputs, y(n) represents the filtered signal, and X represents the filtered output in the frequency domain.

In the **noise removal method** based on **wavelet transforms**, the signal is first altered to the maximum overlapping discrete wavelet domain [57], and the noisy segments are accustomed to a limiting threshold [58]. The soft and hard threshold-based comparison of the signal is carried out so the error lies below a certain limiting value, ε_{ABS}. The decomposition is relaxed only in case the error is higher. The inverse wavelet transform of the maximum overlapping-based threshold signal is carried out. The denoising mechanism using wavelet transforms is illustrated in **Figure** 8, where a is the reconstructed signal with a hard threshold, b is the reconstructed signal with a hard threshold, O indicates the original signal, $\varepsilon(a, O)$ is the error between O and b, $\varepsilon(b, O)$ represents the error between O and b, and $\mathrm{T_h}$ indicates the threshold for the error.

D. LOCALIZATION

The knowledge of the location of the combatant strengthens the planning strategy of a battleground. Gaikwad *et al.* [7] presented a method for identification of source-location based on self-location and self-direction within an artificial battle scenario created by the internet of battlefield things (IoBT). The acoustic source position can be estimated by different approaches such as minimum mean square error and time delay measurements [59]. The estimate of localization or positioning of the acoustic event source is possible by time delay measurements for all nodes with respect to the reference node. The time delay measurement technique for the direction of arrival is given by:

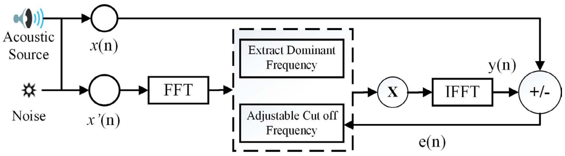


FIGURE 7. Schematic illustration for ABFF.

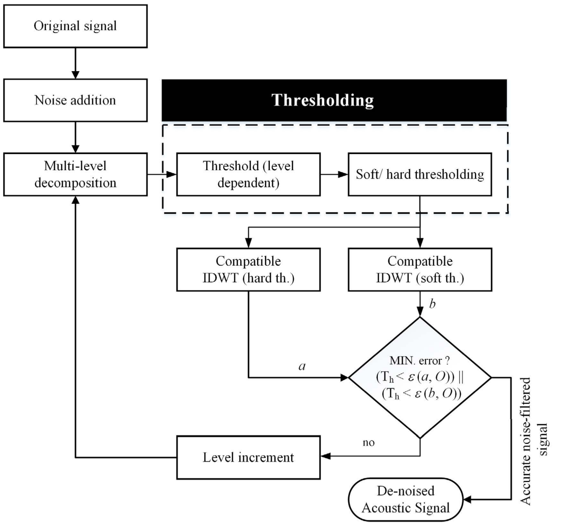

Original signal
Noise addition
Thresholding
Multi-level decomposition
Threshold (level dependent)
Soft/ hard thresholding
Compatible IDWT (hard th.)
Compatible IDWT (soft th.)
b
a
MIN. error ? (T_h < ε(a, O)) || (T_h < ε(b, O))
no
Level increment
De-noised Acoustic Signal
Accurate noise-filtered signal

FIGURE 8. Acoustic signal denoising using Wavelet transforms.

i. Time delay estimation (time domain) is based on finding the peak of the cross-correlation for all nodes (other than the reference node) with respect to the reference node [60].
ii. Time delay estimation (frequency domain) is based on the Hadamard product of acoustic signals transformation to the frequency domain [61]. The cross-correlation of acoustic signals with respect to the reference microphone is determined, and the integral of all time delays is transformed to the time domain.
iii. Similarly, the time delay estimation is carried out using phase transformation. [62].

The measurement of ime delay by any of the techniques follows the computation of the direction of arrival as given by [50]:

$$\tau = -\frac{\boldsymbol{RK}}{c}, \tag{2}$$

where

$$\boldsymbol{K} = \begin{bmatrix} \mathrm{k}_x \\ \mathrm{k}_\mathrm{y} \\ \mathrm{k}_\mathrm{z} \end{bmatrix} = \begin{bmatrix} \sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\phi) \\ \cos(\theta) \end{bmatrix} \text{ and } \boldsymbol{R} = \begin{bmatrix} \overrightarrow{r_2} - \overrightarrow{r_1} \\ \overrightarrow{r_3} - \overrightarrow{r_1} \\ \overrightarrow{r_4} - \overrightarrow{r_1} \end{bmatrix}.$$

Here, τ is the time delay measured as a ratio of the projection of the distance vector, amid microphones along the $\boldsymbol{K}$ direction, to the sound velocity c, and $\boldsymbol{R}$ is the vector based on microphones' distances. The azimuth ϕ and elevation θ angles are representatives of the direction of arrival and can be found using $\boldsymbol{K}$ in 3D Euclidean space as given by:

$$\phi = \tan^{-1}\frac{k_y}{k_x}, \tag{3}$$

$$\theta = \tan^{-1}\frac{k_{xy}}{k_z}, \tag{4}$$

where $k_z = \left(1 - k_{xy}\right)^{\frac{1}{2}}$.

The exploitation of acoustic signal in frequency domain was carried out by [59] and [63]. They dealt with muzzle blast and shock waves separately. Muzzle blast was caused by an eruption of the explosive while a bullet's thrust originates the shock wave. The notion behind the use of frequency domain is that both of these signals are generated at different time domains and hence eventuated in different frequency bands. The signals captured at each microphone are converted into frequency domain with the movable finite-length window to frame the signals in order to find the frame cross-power spectrum of the received signals. Once the signals have been transformed into the frequency domain, phase transformation has been employed to the cross-power spectrum according to certain weights. Finally, the frame cross-correlation function is determined by taking inverse Fourier transformation and τ is determined by the peak value of the cross-correlation that represents the delay τ_{12} between the two signals as given by [64]:

$$R_{12} = \int \psi\left(\omega\right).G_{12}\left(\omega\right)e^{j\omega\tau}d\omega, \tag{5}$$

where R_{12} represents the cross-correlation in the frequency domain between mic-1 and mic-2, G_{12} indicates the cross-power spectrum of the signals r_1 and r_2 that are received at mic-1 and mic-2 respectively. ψ represents the phase transformation of all other mics to get the cross-correlation with reference mic.

The proposed gunshots localization method in windy conditions is described mathematically as follows:

Let $[S_1(n), S_2(n), S_3(n), S_4(n)]$ represent the original gunshot signals and $[N_1(n), N_2(n), N_3(n), N_4(n)]$ denote the wind noise signals at microphones-1,2,3 and 4 respectively.

The first step is to take the Hadamard product of the original gunshot signals with the wind noise signals.

$$x_i(n) = S_i(n) \odot N_i(n) \text{ where, } i = 1, 2, 3, 4$$

The symbol $\odot$ represent the Hadamard product.

The generalized analysis and synthesis sections of discrete wavelet transforms (DWT) has been illustrated in **Figure** 9. Here, the reconstructed $\tilde{x}_i(n)$ can be written mathematically as follows:

$$\tilde{x}_i[n] = \sum_{j=1}^{J}\sum_{k} y_1^j[k]\, g_1^j\left[n - 2^j k\right] + \sum_{k} y_0^j[k]\, g_0^j\left[n - 2^j k\right], \tag{6}$$

where

$$y_1^j[k] = \sum_{n} x_i[n]\, h_1^j\left[2^j k - n\right],$$

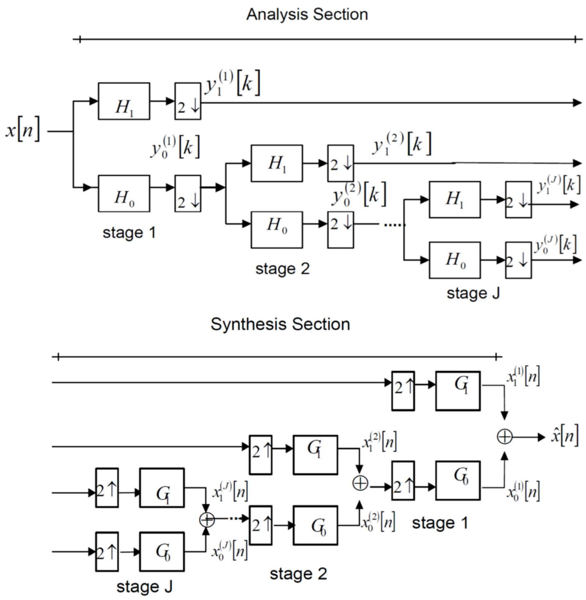


FIGURE 9. The generalized DWT for analysis and synthesis sections using J stages.

and

$$y_0^j[k] = \sum_{n} x_i[n]\, h_0^j\left[2^j k - n\right] \quad j = 1, 2, \ldots, J$$

Now, $a_i(n)$ and $b_i(n)$ are the signals obtained after applying the hard and soft threshold on the reconstructed signal $\tilde{x}_i[n]$ respectively.

The value of J is incremented if the following two conditions are not full filled:

$$T_h < \varepsilon(a, O) = \|\tilde{x}_i[n] - a_i(n)\|^2,$$
$$T_h < \varepsilon(b, O) = \|\tilde{x}_i[n] - b_i(n)\|^2,$$

where $\boldsymbol{T_h}$ is the limiting error value, a is the hard threshold-based reconstructed signal, b is the soft threshold-based reconstructed signal, O is the original signal, $\varepsilon(a, O)$ represents the error between O and a, and $\varepsilon(b, O)$ represents the error between O and b.

The time delay is estimated by finding the peak of the cross-correlation for all the microphones with respect to the reference microphone.

Let

$$\tau = [\tau_{12}, \tau_{13}, \tau_{14}],$$

where $\tau_{12}, \tau_{13}, \tau_{14}$ are the time delays of microphone-2, 3 and 4 with respect to microphone-1, respectively. The ϕ and θ angles can be estimated using **Equations** 2, 3 and 4.

E. CLASSIFICATION

The cognition of the type of the signal is mandatory to acquire the whereabouts of the source because each stimulus falls in a certain bandwidth and has a specific energy spectrum associated with it [65]. Similarly, different classes of guns have distinct acoustic attributes exclusively belonging to them [66].

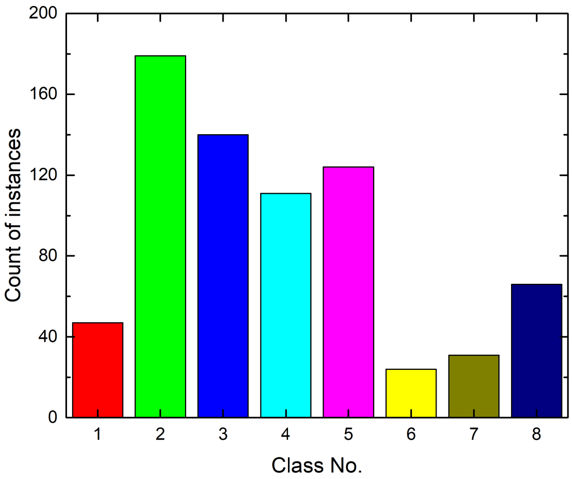


FIGURE 10. Gunshots event distribution across different classes of Airborne dataset (Class No. as annotated in Section 2.1).

Furthermore, it is hard as well as essential to distinguish between an actual gunshot and a non-life-threatening sound disguised as a gunshot signal [67]. We have two sources of acoustic events in this work, one is the Airborne dataset, and the second is the simulated datasets for the original (without noise), normal wind noise, and strong windy conditions for the acoustic event classification problem [68]. The Airborne dataset is based on the gunshots types (**Table** 1). The acoustic event counts vary a lot for different classes across the dataset leading to a poor generalization of the AI-based model. We addressed the class imbalance problem using the synthetic oversampling minority technique (SMOTE).

1) CLASS IMBALANCE

The predictability of the ML algorithm suffers if the number of samples in each class varies during the training phase. The variation of instances is illustrated in **Figure** 10 in the original dataset. In general, the ML model is not up to the mark to learn from acoustic events of minority classes affecting their class (individual) predictive accuracy. To address this issue, it is mandatory to enhance the instances of minority classes. One out of many solutions is SMOTE that improved the number of instances for a class having lesser number of training events. In this technique, the instances of the minority classes are generated by linear interpolation using randomly the k-nearest neighbors of actual events. This improved the generalization of the minority class [69], [70].

2) AUDIO FEATURES-MFCC

The Mel-frequency cepstral coefficients, a frequency group depicting the general shape of the spectral envelope, are used as discriminative audio features for classification. It has been used in the speech recognition arena along with the machine learning paradigm [71]. It is based on extracting the amplitude spectrum of the audio signal in a vector form. Raj *et al.* [72] introduced a multi-layered CNN-based auto-CODEC using MFCC as the input for denoising the signals robustly and securely with an accuracy of 93.25%. Recently, Siam *et al.* [73] used MFCC features obtained from a physiological signal, namely Photoplethysmography (PPG) to find volumetric changes in blood circulation, which is fed to the neural networks to analyze blood flow variation in organs depending on the heartbeats. The PPG signals are reported to be useful as an alternative identification biometric. Unique 13 MFCC [74], [75] are extracted using the acoustic signals of three types, namely original, normal and strong wind noise-impregnated sets. Mel is a short name for melody, and it is based on the comparison of pitch on a scale. The sound recognition systems use these features to identify the speaker [75]. The MFCCs are based on the human ear's perception of the sound frequencies. The human ears receive low frequencies better while the higher tolls are not acknowledged by the receptors. The steps involved in MFCC extraction are given by:

i. First of all, the acoustic signal is segmented so that each frame is 25 milli-second [75]. The number of frames must be an even number (use zero padding in case a frame count is an odd number).
ii. Each segment of the acoustic signal is used to compute the discrete Fourier transform (DFT) as given by: $A_i(k) = \sum_{n=1}^{N} a_i(n) f(n) e^{-j2\pi kn/N}$, $1 \leq k \geq K$, where $a_i(n)$ is i$^{\text{th}}$ frame acoustic signal in the time domain, DFT of i$^{\text{th}}$ frame is $A_i(k)$, the analysis window of the signal is $f(n)$, and K represents the length of the DFT. The spectral density estimate for the segmented acoustic signal is given by: $x_i(k) = |a_i(k)|^2 / N$.
iii. The filter bank (Mel) equates the spectral density to the Mel scale, and a group of 78 standard triangular filters is used for this conversion. In the filter bank, depending on the DFT setting in the previous step, a set of 78 vectors (each of length 257) is chosen. The filter bank energies are computed by summing up the products of each filter bank with spectral density. We have now 78 numbers representing the energy of each filter bank. The computation of log power of 78 Mel frequencies is carried out corresponding to 78 filter bank energies (FBEs).
iv. The cepstral coefficients, 78 in number, are obtained by computing discrete cosine transform (DCT) of 78 FBEs.
v. Only the 39 larger coefficients, reflecting the prompt energy variation in a bank, are selected for the statistics of Minimum, Maximum, Variance and Mean. A vector of these 156 features was considered as Mel frequency cepstral coefficients (MFCC).

Further, GFCC, PNCC, and DWT have also been used as audio features for the classification of acoustic events.

3) OVERVIEW OF ML CLASSIFIERS

We have used many algorithms for spot-checking, and a few algorithms are selected on the basis of computational complexity and prediction results. All of these top-rated machine learning algorithms are experts in their own solution space depending on their parametric fine-tuning according to the structure of the problem. A gist of the classification methods implemented in this work is given by:

An ELM is considered by non-discrete or non-differentiable relations for decreasing the loss function resulting in performance improvement. It has no requirement for parametric optimization. Its basic architecture consists of input and output layers with only one hidden layer. Two types of weight vectors have been used in ELM, one between the input and hidden layers, and the second between the hidden and output layers [69], [76]. Two activation functions are employed in ELM. The sigmoid function is activated amid input and hidden layers, while a linear function is employed amid the hidden and output layers. The objective here is to initialize the input weight vector using a uniform distribution [77]. The ELM output is: $y_k = \sum_{j=1}^{m} \beta_{j,k} g(\sum_{i=1}^{n} w_{i,j} x_i + b_i)$, where $j = 1, 2, 3, \ldots, m$, where $w_{i,j}$ represent input weight vectors, b_i is the bias, x_i is the input feature vector, Y_k represents the output, m indicates the number of neurons in the hidden layer, n represents the number of features in the input layer, and k represents the number of classes in the dataset.

An SVM classifier works on the principle of maximizing the gap between the support vectors, the least confidence points of the classes so that the optimal decision surface divides the instances into their corresponding classes. Depending on the number of classes in the problem, multiple decision surfaces are formed using the training features but the notion is to find the optimum bifurcating surface(s) with maximum margin on either side. The main task is to estimate the weight vector along with bias for classifying all the test instances 'x' with optimum accuracy. The equation for maximum margin is given by:

$$m = 2 * \frac{||w^t x^t + b||}{| |w| |}, \tag{7}$$

where m represents the margin that is twice the gap between the decision hyperplanes, and the support vectors on its either sides. In case the classes are inseparable, the training features are raised to a higher dimension with the help of non-linear kernels. The best-suited kernel and its appropriate parameters can influence the results a lot.

Naïve Bayes (NB) classifier is a probability-based approach where the uncorrelated features of all classes are obtained [78]. The main idea of the NB classifier is derived from the Bayes theorem which states that the posterior probability of an unknown instance can be found in the class distribution of all the classes and their respective class priors are known. Eventually, the event (with an unknown class label) having the larger value of posterior probability for a class would be assigned the same class label. The main assumption of this classifier is that the predictors remain indifferent to the features and the features in an event are independent of one another. NB is based on the posterior probability so that depending on the maximum value corresponds to the selected class. The conditional probability is given by: $P(C_k \mid X_1, X_2, X_3, \ldots, X_n)$ for k outcome or class C_k. The BT is given by:

$$P(C_k \mid X) = \frac{P(C_k)\, P(X \mid C_k)}{P(X)}, \tag{8}$$

where $P(C_k \mid X)$ denotes the posterior probability, $P(C_k)$ is the prior probability, $P(X \mid C_k)$ denotes likelihood and $P(X)$ is the evidence. Naïve Bayes theorem can be expressed as: $P(C_k \mid X_1, X_2, X_3, \ldots, X_n) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(X_i \mid C_k)$, where evidence $Z = P(x)$ is a scaling factor dependent on $(X_1, X_2, X_3, \ldots, X_n)$. The parametric estimation necessitates the use of a marginal probability distribution for the class prior $P(C_k)$ and conditional probability for each known instance given the class $P(X_i \mid C_k)$. The type of x_i (discrete or numeric), such distributions can be multinomial or normal respectively, and it is computed for reach c_j. For inference, the maximum-a-posteriori (MAP) is used, i.e. for $< x_1, x_2, \ldots, x_n >$ select the class c^* such that:

$$PC^* = \underset{c_j}{argmax}\, P\left(C = c_j \mid X_1 = x_1, x_2, \ldots, X_n = x_n\right), \tag{9}$$

$$PC^* = \underset{c_j}{argmax}\, P\left(C = c_j \prod_{i=1}^{n} P(X_i = x_i \mid C = c_j)\right). \tag{10}$$

NB is a simple and unrealistic independence method and has efficiently been used in many applications.

A k-NN classifier, a non-parametric algorithm, is a proximity-based approach where the number of nearest neighbors of the unknown event, in the feature space, decides the class label assignment to the unknown event[79]. The notion behind its working is that similar things are contiguous to each other [80]. The k-NN use for a test sample classification is based on finding its distance from all other training instances. The metrics that can be used for distance measurements can be Hamming, Euclidean, Manhattan, or any other distance measuring scheme. The paths or distances are then sorted in an ascending order to track the nearest neighbor listing. First k number of nearest neighbors are selected, and each one of the points causing the shortest distance then votes for the unknown event. The class having the maximum number of votes for the unknown event is used to assign its class label. The value of 'k' plays a vital role in the working of k-NN and it mostly depends on the nature of the underlying data distribution.

RF is an ensemble-learning classifier that uses manifold decision trees (DTs) as the base learners [81]. Each one of the DTs is designed on the basis of random attribute selection from training instances. After the training of all the decision trees, the test event is passed through each one of them, and the DTs will opt for the class in terms of a vote for each test event. The class having the maximum number of votes will be assigned to the test event.

4) 1D-CNN ARCHITECTURES FOR MULTI-CLASS CLASSIFICATION

We have introduced three 1D-CNN architectures based on the number of layers, namely Light, Mild and Extensive 1D-CNN architectures as illustrated in **Figure** 11. The number of

TABLE 2. Parameters used for 1D-CNN architecture used for classification of gunshot acoustic events.

Parameters	Values
Learning rate	0.0005
Epoch	100
Performance evaluation threshold	37
Patience Level	10
Total Layers	12
Kernel Size	1,3
Pool Type	MaxPooling 1 × 2
Input Size	156 × 1
Drop out	0.25
Solver Name	Adam
Batch Size	2

layers varied to extract the features in the encoding part of the architectures. **Figure** 11 (a, b, & c) illustrates the buildup of three architectures using 10, 12, and 18 layers respectively.

Each of the 1D-CNN architectures is based on 1D-convolution between the 1D-receptive length of the signal with an associated set of kernels to extract the features dynamically based on the general working principles of deep learning. Here, we have used MFCC feature vectors from the acoustic signal data. The sizes of layers that have been used are depicted in the design of the respective architecture. The set of parameters used to extract features by 1D-representation learning is illustrated in **Table** II.

F. LOCALIZATION AND CLASSIFICATION PERFORMANCE MEASURES

For noise analysis, the variations are in the form of filter parameters along with $(\mathrm{SNR})_{\mathrm{Noise}}$. The subscript "Noise" has been added to indicate the higher noise content in the ratio. The $(\mathrm{SNR})_{\mathrm{Rec}}$ has been used for the estimation of the quality of reconstructed signals as compared with the original signals in decibels [82]. The subscript "Rec" has been added to indicate the higher reconstructed signal content in the ratio. The good quality of the restored signal is indicated by the higher $(\mathrm{SNR})_{\mathrm{Rec}}$ value. The mathematical relation for the signal-to-noise ratio is given by $SNR = 20.\log_{10}\left(\frac{S}{N}\right)$, where S represents the desired signal level and N represents the noise signal level [83], [84].

The performance measurement of the localization phase to specify the position of the acoustic even source in 3D-space is carried out using the (φ, θ) angles pair. The localization performance is computed by means of relative error as given by [85]:

$$E_{loc} = \frac{\left|\left|e_m - e_p\right|\right|}{e_p}, \tag{11}$$

where e_m and e_p represent the measured and actual positions of the acoustic source. The localization performance as accuracy is given by: $A_{loc} = (1 - E_{loc}) * 100$.

For classification performance measurement, one of the important measures for the ML model is the classification accuracy as given by [86]:

$$A_{cls} = \frac{T_P + T_n}{T_p + T_n + F_P + F_n}, \tag{12}$$

where

T_p = true positives: correctly identified acoustic events by the model,

T_n = true negatives: correctly identified events as negatives (the other class even correctly classified in comparison to the positive class) by the model,

F_p = false positives: incorrectly classified the negative samples as of positive class,

F_n = false negatives: incorrectly classified positive samples as that of negative class.

Another performance measure is the specificity that is used to evaluate the model by using a false-positive rate as given by [86]:

$$S_p = \frac{T_n}{F_p + T_n} \tag{13}$$

Similarly, the performance measure recall (or sensitivity) is used to evaluate the model by using the true positive rate as given by [80]:

$$R_c = \frac{T_p}{T_p + F_n} \tag{14}$$

Precision is another performance measure to check the positive class predictions (T_p, F_p) that are actually from the positive class as given by [80]:

$$P_r = \frac{T_p}{T_p + F_p} \tag{15}$$

Another popular classification performance measure is F-score which is considered most effective for imbalanced datasets. It is defined as the harmonic mean between P_r and R_c as given by [86]:

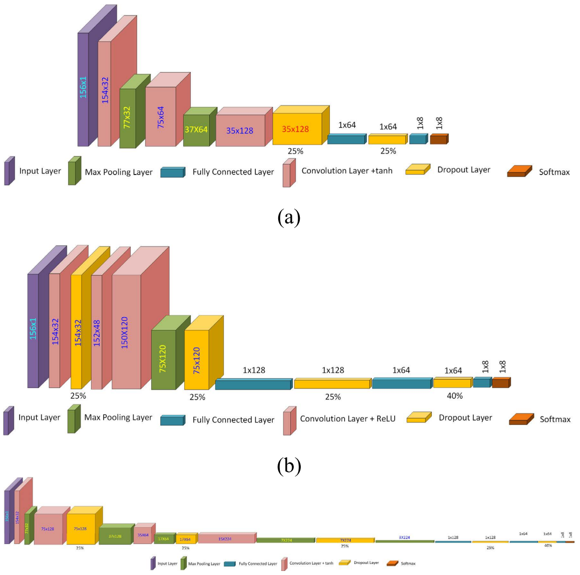
$$F - score = \frac{2 \times P_r \times R_c}{P_r + R_c} \tag{16}$$

III. RESULTS AND DISCUSSION

The entire experimentation was carried out using a computer machine having A6-6310 processor, AMD Radeon R4 graphics card, and 16 GB RAM modules. The open source libraries were used to develop scripts for localization and classification tasks. For machine learning models, the training: test ratio used is 90:10 with stratified resampling.

A. FEATURES' VISUALIZATION

The enormous data generated in most scientific fields make it challenging to visualize it. In our case, to visualize MFCC distribution in feature space, we carry out the mapping from high to low dimensional feature space by adopting the t- Stochastic Neighbor Embedding (t-SNE) plot. Here the high dimensional points are mapped to a low dimensional space so that the former behavior is mimicked by the latter one. The algorithm used to carry out this task model uses two



(a)

(b)

(c)

FIGURE 11. **CNN architectures used for acoustic dataset: (a) Light 1D-CNN architecture, 10 layers, (b) Mild 1D-CNN architecture, 12 layers, and (c) Extensive 1D-CNN architecture, 18 layers.**

distributions: original points (Gaussian distribution based) and the embedded points (Student's t-distribution based) [87]. The divergence (Kullback-Leibler) between these distributions is minimized by positional variation of the embedded points [88]. The event distribution in classes is illustrated in **Figure** 12. The challenging nature of classification using this dataset is evident due to the overlapping class boundaries of the acoustic events in the feature space.

B. NOISE STUDY

The filtering of noise is affected by changing $(\mathrm{SNR})_{\mathrm{Noise}}$ after the simulation of sound events with models of normal wind and a strong windy environment. In real scenarios, $(\mathrm{SNR})_{\mathrm{Noise}}$ goes up to -20 dB in high noise levels, but for wind sensitivity analysis filters work if the noise level decreases up to -3100 dB although this is not a real scenario. But parameters of both adaptive and conventional filters are observed for the best acoustic detection as illustrated and discussed in the subsequent tables.

The sole objective of this part of experimentation is to critically scrutinize the performance of filters by empirical parametrization. The results of conventional filter experimentations are shown in **Tables** III & IV for normal and strong windy models respectively. During this experimentation, the $(\mathrm{SNR})_{\mathrm{Noise}}$ value was kept at -3100 dB. It was unrealistic but it was adopted in order to check the high-value response of $(\mathrm{SNR})_{\mathrm{Noise}}$ for our acoustic signals. It was observed that raise in noise level resulted in the alleviation of filter performance. This suppression caused by noise could be parametrically controlled by window width (W_w) and the polynomial order (P_o). The optimization of parameters was carried out by keeping one parameter constant while modifying the other. It was found that the Savitzky-Golay filter with ($W_w = 27$ & $P_o = 4$) achieved an $(\mathrm{SNR})_{\mathrm{Rec}}$ of 60.36 dB for the normal wind noise model. In the same acoustic conditions, the median filter with ($W_w = 5$ & $P_o = 1$) achieved $(\mathrm{SNR})_{\mathrm{Rec}}$ as 69.82 dB. The high value of $(\mathrm{SNR})_{\mathrm{Rec}}$ using the median filter is attributed to its suppression of noise peaks. Further, the comparison of Savitzky-Golay and median filters shows that the complexity of median filters is low due to the reduced number of frame sequences. The moving average filter performance was not up to the mark reconstruction of the acoustic signal due to the noise averaging effect. The conventional filter experimentation for strong wind noise showed that the same trend followed as that of normal wind noise. The results, however, suffered as illustrated in **Table** IV due to extreme conditions of wind noise. For strong wind model based acoustic signal filtering, the Savitzky-Golay filter resulted in an $(\mathrm{SNR})_{\mathrm{Rec}}$ of 51.44 dB which was lower than the best performing median filter with an $(\mathrm{SNR})_{\mathrm{Rec}}$ of 61.58 dB. Similarly, the strong windy environment-based performance of the Savitzky-Golay filter (51.44 dB) can be compared with a normal wind environment (60.36 dB), thereby validating the effect of wind-noise extreme conditions evident.

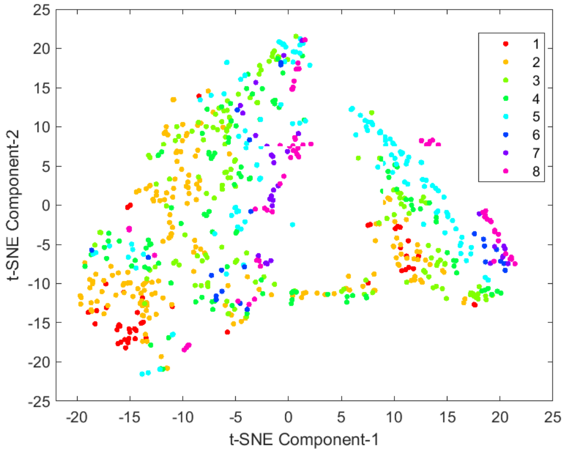


FIGURE 12. **The feature set visualization of acoustic events extracted from Airborne dataset using t-SNE plot.**

Similarly, the experimentation using adaptive filters resulted in **Tables** V & VI with empirically found parameters illustrating the thresholds of $(\mathrm{SNR})_{\mathrm{Noise}}$ for normal as well as strong windy noise models respectively. The performance of adaptive filters has been found outclass in comparison to the conventional filters. Further, the wavelet transforms with Daubechies (db) family (soft threshold at level 1) resulted in an outclass performance on a noisy signal, $(\mathrm{SNR})_{\mathrm{Noise}} = -3100$ dB, achieving $(\mathrm{SNR})_{\mathrm{Rec}} = 178.26$ dB under a normal windy environment. The application of the same set of parameters for wavelet de-noising filter using a strong noisy signal resulted in a degraded signal (94.26 dB).

TABLE 3. Conventional filters analysis for normal wind noise (W_W and P_O represent window width and polynomial order respectively, $(SNR)_{Noise}$ is the noise impregnation level and $(SNR)_{Rec}$ is the reconstructed signal quality).

Filter	$(SNR)_{Noise}$ dB	Parameter		$(SNR)_{Rec}$ dB
		W_w	P_o	
Savitzky-Golay	-3100	27	4	60.36
Median	**-3100**	**5**	1	**69.82**
Moving Average	-3000	9	1	32.87

TABLE 4. Conventional filters analysis for strong wind noise (W_W and P_O represent window width and polynomial order respectively, $(SNR)_{Noise}$ is the noise impregnation level and $(SNR)_{Rec}$ is the reconstructed signal quality).

Filter	$(SNR)_{Noise}$ dB	Parameter		$(SNR)_{Rec}$ dB
		W_w	P_o	
Savitzky-Golay	-3100	27	4	51.44
Median	**-3100**	**4**	1	**61.58**
Moving Average	-3000	9	1	28.18

TABLE 5. Adaptive filters analysis for normal wind noise (WT stands for wavelets transforms, db1 represents Daubechies family with a soft threshold at level 1, $(SNR)_{Noise}$ is the noise impregnation level and $(SNR)_{Rec}$ is the reconstructed signal quality).

Filter	$(SNR)_{Noise}$ dB	Parameter	$(SNR)_{Rec}$ dB
LMS	-1000	Step size = 0.354 Order = 1	77.732
RMS	-3100	Step Size = 0.001	100.47
WT	**-3100**	**db1**	**178.26**
ABFF	-2900	Cut off : 100 MHz AC: 60	82.83

TABLE 6. Adaptive filters analysis for strong wind noise (WT stands for Wavelets Transform, db1 represents Daubechies family with a soft threshold at level 1, $(SNR)_{Noise}$ is the Noise Impregnation Level and $(SNR)_{Rec}$ is the Reconstructed signal Quality).

Filters	$(SNR)_{Noise}$ dB	Parameter	$(SNR)_{Rec}$ dB
LMS	-1000	Step size μ= 0.354 Order = 1	71.732
RMS	-3100	Step Size μ = 0.001	87.11
WT	**-3100**	**db1**	**94.26**
ABFF	-2900	Cut off: 100 MHz AC: 60	79.83

C. LOCALIZATION STUDY

The original sound events were simulated using four distant microphones (70 cm) arranged in an omnidirectional geometry (**Section** 2.2) keeping one microphone serving as the reference point. The simulated signals are then subjected to noise addition at different noise levels. We investigated different distances from the reference microphone (viz. 50 cm, 70 cm, and 100 cm), and the actual and localization accuracy was found the same as for 70 cm, rather than the computational cost that appeared as a consequence.

The reference microphone was positioned with known φ and θ relative to the acoustic source measuring 20° and 90° respectively. We assumed the that sound event source was in a windy noise situation. The noise level was added with a known $(SNR)_{Noise}$ at each node thereby lowering the acoustic signal quality. These signals were used for the analysis of localization-based experimentation for time delay estimation. The signals are filtered and reconstructed using conventional and adaptive filters before finally being used for time delay estimation. Consequently, the experimentation was carried out for direction of arrival (φ', θ') estimation (estimated azimuth and elevations angles) along with measuring the quality of the reconstructed signal in terms of $(SNR)_{Rec}$. It has been found that the quality of the acoustic signal is important as it is directly correlated with the localization accuracy. The greater the strength of the acoustic signal, the higher the localization accuracy resulting thereof. The localization precision has been taken care of by considering the results to four decimal places. Further, the experimentation using Hadamard product has been found to perform remarkable, in normal as well as extreme wind conditions, as compared with the other two techniques. The conventional filters have been found sensitive to noise. Consequently, good reconstructed signals after filtering are desirable, using conventional filters, for better localization of the acoustic source.

The localization performance is illustrated in **Tables** VII & VIII for normal and strong windy noise using conventional filters for reconstructed signals. The signals from each microphone are noise impregnated at specific $(SNR)_{Noise}$ to degrade the acoustic signal quality. The various noise levels $(SNR)_{Noise}$ are used to generate the noisy acoustic signals in each of the microphones. The $(SNR)_{Noise}$ at microphone-1 (reference microphone) was kept at 14 dB assuming a high-quality signal of the acoustic source. Similarly, $(SNR)_{Noise}$ was kept -2dB at microphone-2 considering more noise affected signal, while microphone-3 was kept at

TABLE 7. **Localization study using conventional filters under normal wind conditions (Sim. No. represents the Experimentation number, and Mic. represents the microphone, $\varnothing'$ and θ' represent the estimated azimuth and elevation angles respectively, and $(SNR)_{Rec}$ is the Reconstructed signal Quality).**

Sim. No.	Filters	Mic. No	$(SNR)_{Rec}$ dB	$\varnothing'$	θ'	A_{loc} (%)
1	Moving Average	1*	16.63	20.0699	90.4204	99.95
		2	8.09			
		3	11.33			
		4	8.05			
2	Median	1*	24.59	20.6563	90.2099	99.95
		2	8.65			
		3	12.58			
		4	8.58			
3	Savitzky-Golay	1*	17.37	20.0699	90.4204	99.95
		2	8.19			
		3	11.54			
		4	8.17			

* Mic. No 1 represents the reference node.

TABLE 8. **Localization study using conventional filters under strong wind conditions (Sim. No. represents the Experimentation number, and Mic. represents the microphone, $\varnothing'$ and θ' represent the estimated azimuth and elevation angles respectively, and $(SNR)_{Rec}$ is the Reconstructed signal Quality).**

Sim. No:	Filter	Mic. No	$(SNR)_{Rec}$ dB	$\varnothing'$	θ'	A_{loc} (%)
1	Moving Average	1*	16.56	19.6563	90.2099	99.95
		2	8.08			
		3	11.37			
		4	8.11			
2	Median	1*	24.61	19.6563	90.2099	99.95
		2	8.64			
		3	12.57			
		4	8.62			
3	Savitzky-Golay	1*	17.35	19.6563	90.2099	99.95
		2	8.19			
		3	11.55			
		4	8.16			

* Mic. No 1 represents the reference node.

2 dB considering the noise impregnation level between the first and third microphones. The $(SNR)_{Noise}$ was kept at -2 dB at microphone-4 allowing for weak power to receive acoustic signals. **Table** VII illustrates the localization performance by using the Hadamard product-based time delay estimation using reconstructed signals from conventional filters under normal wind conditions. The median filter-based localization accuracy has been found to be 99.95% using the median filter (in the conventional filters group) which has been found to be outclass in its class of filters. Although the localization accuracy was computed and found the same for different filters, the azimuth and elevation angles are marginally different. The slight difference in these measurements is important, especially in the case of surveillance systems. It was found that with constant $(SNR)_{Noise}$-based simulation using strong windy conditions, the resulting localization performance was up to 99.95% as illustrated in **Table** VIII.

Moreover, the localization results using reconstructed signals, obtained from adaptive filters, are illustrated in **Tables** IX & X. In these tables, the quality of reconstructed signals in terms of $(SNR)_{Rec}$ is alleviated in the windy environment at a specific noise contamination $(SNR)_{Noise}$. The various noise levels $(SNR)_{Noise}$ are used to generate the noisy acoustic signals in each of the microphones or nodes. The node-1 was kept at $(SNR)_{Noise} = 9$ dB assuming the good quality of the acoustic signal at the reference node, whereas the node-2 was kept at noise level $(SNR)_{Noise} = -2$ dB assuming a severely damaged signal. However, at node-3 noise level was kept at $(SNR)_{Noise} = 1$ dB assuming signal quality between node-1&3. At node-4 the noise level was $(SNR)_{Noise} = -4$ dB assuming the lower capturing quality of the microphone along with the noise contamination effect. **Table** IX shows the localization performance of the Hadamard product using the reconstructed signals from adaptive filters under normal windy conditions. The localization performance accuracy has been found to be 99.98% with wavelet denoising filters on reconstructed signals. It was further observed that in the signal simulation based on specific $(SNR)_{Noise}$ under extreme wind conditions, as illustrated in **Table** X, the localization accuracy was dropped up to 99.95% using the wavelet transforms-based denoising method showing its consistent behavior under varying wind noise conditions.

The performance of localization on reconstructing signals obtained from ABFF is accurate up to 98.03% under normal windy environments. When compared with wavelet denoising filters the ABFF filtering capability increases in strong wind noise impregnation to the original signal but its localization capability is less precise. This may be attributed to the fact that in the presence of strong wind noise the amplitude of the acoustic source is distinct and identifiable easily. The localization performance using RLS and LMS has similar behavior. However, in case the reconstructed signals are based on LMS in strong wind noise, the localization performance is reduced to 99.91%, whereas the RLS has shown stable behavior in strong wind noise due to recursion algorithms.

The filtering is expected to leave behind the desired features of a gunshot signal, which often has noise-like constituents, however, the accuracy of our results shows that the effect of filtering has been found negligible as shown in **Tables** VII-X.

D. CLASSIFICATION PERFORMANCE ANALYSIS OF PROPOSED TECHNIQUE USING MFCC

The multi-dimension feature extraction is carried out using MFCC statistics based on min, max, mean and standard deviation. The MFCC has 156 coefficients (Figure 12) indicating the dimensionality of the problem. We employed ML algorithms for ideal (original signals) and noisy acoustic environments as illustrated in **Tables** XI-XIV. In the case of ELM, an empirical analysis of the number of neurons in the

TABLE 9. Localization study using adaptive filters under normal wind conditions (Sim. No. represents the Experimentation number, WT stands for Wavelets Transform, and Mic. represents the microphone, $\emptyset'$ and θ' represent the estimated azimuth and elevation angles respectively, and $(\mathrm{SNR})_{\mathrm{Rec}}$ is the reconstructed signal quality).

Sim No:	Filter	Mic No	$(\mathrm{SNR})_{\mathrm{Rec}}$ dB	$\emptyset'$	θ'	A_{loc} (%)
1	LMS	1* 2 3 4	25.49 16.58 18.98 15.77	19.6532	90.2099	99.95
2	RLS	1* 2 3 4	20.80 12.40 14.33 11.72	19.6541	90.2099	99.95
3	ABFF	1* 2 3 4	11.19 8.86 9.85 8.85	43.8912	76.4938	96.80
4	WT	1* 2 3 4	20.38 8.62 12.53 6.58	19.7971	90.0000	99.98

* Mic. No 1 represents the reference node.

TABLE 10. Localization study using adaptive filters under strong wind conditions (Sim. No. represents the experimentation number, WT stands for wavelets transform, and Mic. represents the microphone, $\emptyset'$ and θ' represent the estimated azimuth and elevation angles respectively, and $(\mathrm{SNR})_{\mathrm{Rec}}$ is the reconstructed signal quality).

Sim. No.	Filters	Mic. No.	$(\mathrm{SNR})_{\mathrm{Rec}}$ dB	$\emptyset'$	θ'	A_{loc}
1	LMS	1* 2 3 4	23.73 16.42 18.17 15.34	19.2438	90.0000	99.9120
2	RLS	1* 2 3 4	20.09 12.78 14.13 12.04	19.6563	90.2099	99.9532
3	ABFF	1* 2 3 4	11.29 9.89 10.29 8.67	22.6888	80.9616	98.9031
4	WT	1* 2 3 4	20.40 8.60 12.53 6.55	19.6563	90.2099	99.9532

* Mic. No 1 represents the reference node

TABLE 11. Empirical analysis for ELM parameterization under normal windy conditions ($(\mathrm{SNR})_{\mathrm{Noise}} = 20$ dB); N_h represents the hidden neurons, and 55500 neurons that correspond to the optimum accuracy value.

N_h	A_{cls} (%)	S_p	P_r	R_c	F-score
10	0.20	0.08	0.12	0.94	0.22
100	48.95	0.46	0.15	0.67	0.25
1000	71.33	0.71	0.26	0.72	0.38
10000	83.91	0.83	0.43	0.88	0.58
20000	91.61	0.92	0.64	1	0.78
30000	88.81	0.872	0.53	1	0.69
40000	88.12	0.86	0.51	1	0.68
50000	88.12	0.86	0.51	1	0.68
55000	91.61	0.90	0.6	1	0.75
60000	90.05	0.86	0.59	1	0.73
70000	89.51	0.88	0.56	1	0.70
80000	88.81	0.87	0.53	1	0.70
90000	90.10	0.89	0.58	1	0.73
100000	87.41	0.85	0.5	1	0.67

hidden layer, as illustrated in **Table** XI, has been carried out and the model was fine-tuned for our gunshot acoustic signals framework. The activation functions that have been checked include radial basis, triangular basis, sigmoid, sine, hardlim, etc. The sigmoid function was selected with 55500 neurons in the hidden layer as found optimum in **Table** XI. The optimum results were found using ELM with sequential degradation in performance observed in datasets for original acoustic data without noise addition, normal wind, and strong windy environments with $(\mathrm{SNR})_{\mathrm{Noise}} = 20$ dB as 93.01% (**Table** XII), 91.61% (**Table** XIII), and 88.11% (**Table** IV) respectively.

The remarkable classification performance of ELM was observed owing to the fact its ability to find the isolation surface between classes having complex feature spaces with overlapped events across classes using only the discriminant features.

The notion of SVM for non-separable data in the feature space is to raise the dimension of data until it becomes separable by a hyperplane. In our acoustic signal classification problem, the features are raised to a higher dimensional space with a polynomial kernel (order = 2) so that the features become linearly separable. The maximization of margin with this setup resulted in a classification accuracy of 90.91% as shown in **Table** XII. The classification performance dropped sequentially with mild to severe wind noise conditions. The SVM accuracy dropped to 86.11% (**Table** XIII) and 77.08% (**Table** XIV) for normal and strong wind noise models respectively with $(\mathrm{SNR})_{\mathrm{Noise}} = 20$ dB.

The decision boundary formation of the NB classifier depends on the posterior probabilities of acoustic events' classes where the features are assumed to be conditionally independent of one another. As a matter of fact, for explosives data, the features are slightly correlated causing feature overlapping, and this resulted in deteriorated performance. A slight improvement in results is observable in the case of extreme wind noise conditions. Here two competing processes may be thought of as taking place simultaneously, one successfully discriminating the acoustic events with assumptions and the second causing overfitting in the presence of wind noise. The latter superseded the former after a sufficient number of training instances were used in the training phase.

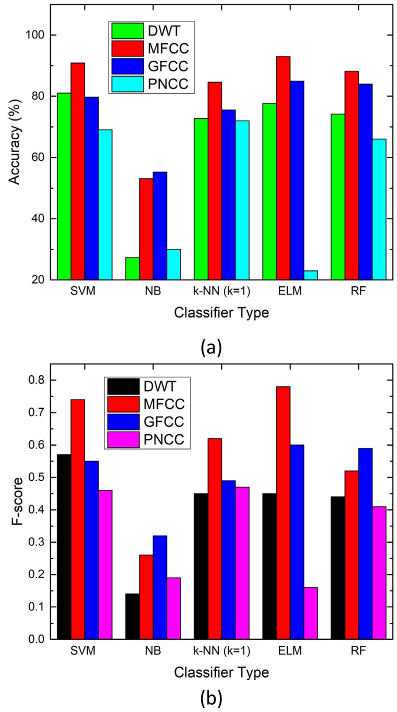
A robust and consistent learning pattern was observed for the RF classifier even in strong windy conditions as shown in **Tables** XII-XIV. In this case, the notion is to average DTs of multiple depths, same trees subjected to different portions of the training partition thereby increasing diversity and alleviating correlation between features. This results in consistent performance in normal as well as extreme wind noise conditions measuring accuracy in the case of RF above 80%. On a similar basis, the k-NN classifier performed relatively better for $k = 1$ due to the correlated features found with a single contiguous instance in good association with the test event. As far as the classification rate in normal and strong windy conditions is concerned, the k-NN classifier has been found reliable and robust. However, in the strong windy environment, there is slight improvement due to an extensive range of feature boundaries with a higher probability of selecting the correct label instance in close proximity to the test instance.

E. CLASSIFICATION PERFORMANCE ANALYSIS OF INDIVIDUAL FEATURE EXTRACTION TECHNIQUES

We investigated DWT [34], MFCC [30], GFCC [28] and PNCC [31], [89] as the potential feature extraction techniques. The analysis of individual feature extraction techniques, viz. DWT, MFCC, GFCC, and PNCC, have been plotted for classification performance as depicted in **Figure** 13. We experimented with SVM, NB, k-NN, ELM, and RF with the feature sets for DWT, MFCC, GFCC, and PNCC. The overlapping features in GFCC, DWT, and PNCC including their hybrids resulted in adverse classification performance. Further, it has been found that the ELM using MFCC for wind noise-contaminated gunshots signals appear the most significant for classification purposes, resulting in relatively outclass performance. **Figure** 13 (a) shows the individual feature type performance in terms of accuracy for classifiers, viz. SVM (polynomial order 2), NB, k-NN ($k = 1$), ELM (55500 neurons), and RF. The overall performance, keeping in view the feature extraction type, has been found to be excellent using ELM with MFCC features. Similarly, as illustrated in **Figure** 13 (b), the F-score has been found high relatively for ELM using MFCCs.

F. CLASSIFICATION PERFORMANCE ANALYSIS OF HYBRID FEATURE EXTRACTION TECHNIQUES

We used a manifold combinatorial logic of two feature types using individual features extracted through DWT, MFCC, GFCC, and PNCC. It has been found using **Figure** 13 that the MFCC boosted the features more as compared with other feature extraction techniques, and resulted in an outclass performance whatever may be the type of classifier. Similarly, it is further inferred that the ELM exploited the MFCC features more as compared with other classifiers. The results of the hybrid of two feature types are illustrated in **Figure** 14. However, not even fractional enhancement in performance was observed with the formation of hybrids like DWT-MFCC [35], GFCC-MFCC [36], and MFCC-PNCC [36]. The results of the individual MFCC features for the Airborne dataset gave the optimum performance as compared with other potential feature extraction techniques.



(a)

(b)

FIGURE 13. Feature type analysis of DWT, MFCC, GFCC, and PNCC techniques on an individual basis for computation of (a) accuracy, and (b) F-score using classifiers: SVM, NB, k-NN, ELM, and RF for Airborne dataset.

Each classifier is accompanied by its possible solution domain, using discriminative features, culminating in performance evaluation measures. When these features are merged, they collectively may or may not generate better results. When the features from different feature sets are combined, if uncorrelated, they can improve discrimination improving performance. The converse is also true. In our acoustic classification problem using the gunshots dataset, SVM, RF, and k-NN classifiers resulted in relatively better performance as compared with ELM in the hybrid of feature types on a classification basis.

G. CLASSIFICATION PERFORMANCE ANALYSIS USING 1D CNN ARCHITECTURES

The experimentation for CNN architectures was based on DWT [33], [34], MFCC [30], GFCC [28], [29] and PNCC [31], [89] feature sets. The results of MFCC were found outclass and the entire section reports the results using MFCC architecture.

It is customary to evaluate the sensitivity of various combinations of fine-tuned architectures in case the performance classification is to be optimized. Three architectures, light, mild and extensive, resulted in performance evaluation as illustrated in **Figure** 15 for original sound events, normal wind noise model, and strong wind noise model respectively. The parametric set used to run the analysis of the code is given in **Section** II. G.

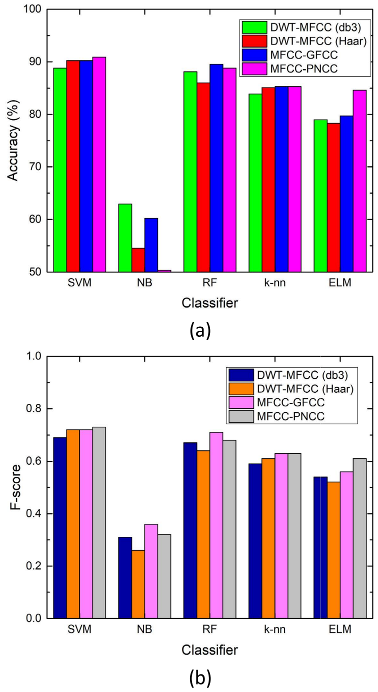


FIGURE 14. Hybrid Feature type analysis of DWT, MFCC, GFCC, and PNCC techniques for computation of (a) accuracy, and (b) F-score using classifiers: SVM, NB, k-NN, ELM, and RF for Airborne dataset.

It is customary to evaluate the sensitivity of various combinations of fine-tuned architectures in case the performance classification is to be optimized. Three architectures, light, mild and extensive, resulted in performance evaluation as illustrated in **Figure** 15 for original sound events, normal wind noise model, and strong wind noise model respectively. The parametric set used to run the analysis of the code is given in **Section** II. G.

Figure 15 (a) shows the improved results with mild 1D-CNN architecture compared with other options tried for a multi-class acoustic classification problem structure. Further, the addition of severity of wind noise in original acoustic signals also resulted in degradation of the classification accuracy. Next to the mild architecture, light architecture resulted in an excellent performance. The extra layers in extensive architecture were not found competing with the lighter architecture options with a relatively lesser number of layers. Similarly, **Figure** 15 (b) shows the comparison of the F-score, which is important for the performance analysis of imbalanced classes. The same trend was found as encountered in the previous case that the mild 1D-CNN architecture resulted in a relatively higher F-score, with the best results found in the untreated noise model case. The addition of noise resulted in the degradation of the signals.

The comparison of classification performance of CNN architectures with ELM showed that the latter produced outclass performance. The difference in efficiency is attributed to the lesser number of instances available during the training phase even with instance augmentation application using SMOTE technique. Deep learning, in general, requires a large number of training instances in which case the results of classification problems suffer drastically otherwise. The deep learning strategy, in this particular gunshot detection problem, was not found as the ultimate solution as it was expected to be.

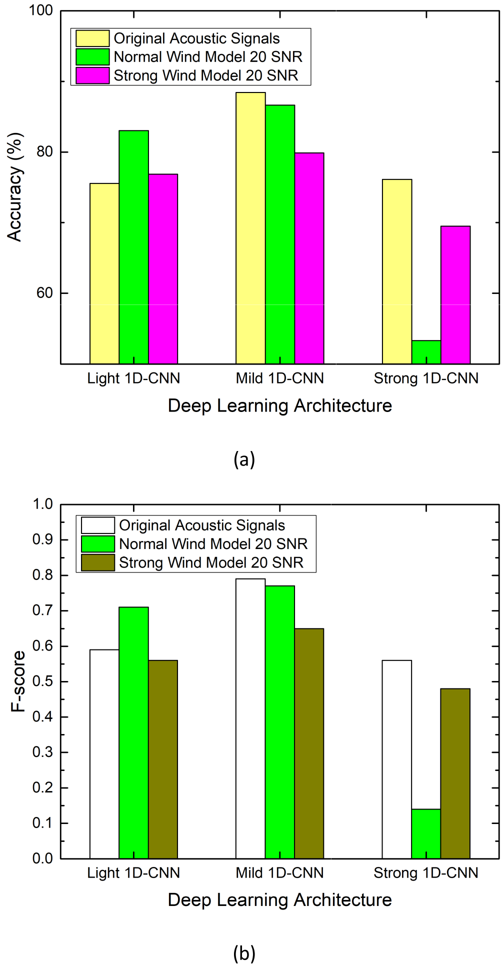


FIGURE 15. The three proposed 1D-CNN architectures for classification performance analysis in terms of (a) accuracy, (b) F-score for original acoustic events, normal and strong wind noise models.

H. COMPLEXITY ANALYSIS OF PROPOSED CLASSIFICATION MODEL

It is a measure of the time complexity of the proposed framework and is attributed to the assessment of resources (time and space) that a specific algorithm uses while executing it on a computer system.

In the tests conducted we took into consideration the components to carry out the complexity analysis of the proposed system of classification as illustrated in **Table** XV. The uniform basis has been selected with the optimum parameters already used in driving the optimum classification results.

The overall reduced prediction time with outclass accuracy per instance for gunshot type classification is achieved using

TABLE 12. Original acoustic signals based comparison of performance.

Classifier	A_{cls} (%)	S_p	P_r	R_c	F-score
ELM (N_h: 55500)	93.01	0.92	0.64	1.00	0.78
SVM (poly order: 2)	90.91	0.90	0.58	1.00	0.74
NB	53.14	0.51	0.16	0.67	0.26
k-NN (k=1)	84.62	0.82	0.45	1.00	0.62
RF	88.81	0.87	0.52	1.00	0.69

TABLE 13. Normal wind impregnated signals based comparison of performance.

Classifiers	A_{cls} (%)	S_p	P_r	R_c	F-score
ELM (N_h: 55500)	91.61	0.90	0.60	1.00	0.75
SVM (poly order: 2)	86.11	0.85	0.39	1.00	0.57
NB	63.89	0.60	0.20	1.00	0.33
k-NN (k=1)	84.72	0.83	0.45	1.00	0.62
RF	85.32	0.84	0.46	0.94	0.62

TABLE 14. Strong wind impregnated signals based comparison of performance.

Classifier	A_{cls} (%)	S_p	P_r	R_c	F-score
ELM (N_h: 55500)	88.11	0.86	0.51	1	0.68
SVM (poly order: 2)	77.08	0.75	0.35	0.94	0.51
NB	60.14	0.56	0.22	0.89	0.36
k-NN (k=1)	86.81	0.85	0.41	1.00	0.58
RF	83.22	0.82	0.42	0.89	0.57

ELM. The comparison of computational times for different potential classifiers during training and testing phases is illustrated in **Figure** 16.

I. COMPARISON WITH EXISTING METHODS

We have compared our work for localization and classification performance using state-of-the-art methods. The selection was made for recently used existing methods, and the implementation was carried out using the Airborne dataset.

The comparison of our localization results with existing methods is shown in **Table** XVI. The classification results for comparison with existing methods are illustrated in **Table** XVII. The hybrids of different individual acoustic feature extraction techniques have also been tried. The hybrid of MFCC-PNCC with SVM (polynomial of order 2) has been found remarkable (90.91% as accuracy), and next to our proposed ELM-based technique (93.01% as accuracy). The lesser number of instances available in the dataset resulted in relatively degraded classification performance of CNN (88.42%).

TABLE 15. Complexity analysis variables/parameters for the proposed classification system.

Variables/ parameters	Value
Total instances (without augmentation)	722
Instances with augmentation	1296
Total augmentation time	514.51 ms
Augmentation time per instance	0.72 ms
Total training instances (80%)	578
Total testing instances (20%)	144

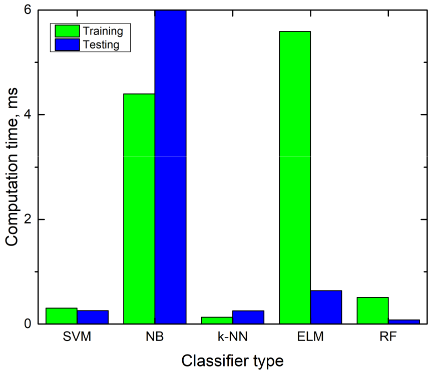


FIGURE 16. Complexity analysis of the proposed system for different classifiers used for Airborne dataset using augmentation by SMOTE.

J. CHALLENGES AND FUTURE RECOMMENDATIONS

The sound effects library used by us does not include the real difference between a close recording at a shooting range vs. a real recording at a considerable distance from the firearm. Similarly, actual gunshot detection systems do not have a pre-recorded single sound, but have to cope with the fact that gunshots are highly directional, have multi-path reflections from the ground and other nearby surfaces, suffer diffraction due to obstacles between the shooting position and the microphones, and vary substantially according to the orientation of the gun's muzzle concerning the microphones. In the future, the real-time recordings of actual gunfire with varying distances, and varying acoustical surroundings from the microphones will be useful to find more realistic localization and range finding of gunshots.

The time-delay detection of a signal depends upon identifying a specific feature that is recognized in each microphone signal. In the presence of noise, there is a considerable likelihood that distant gunshot sounds may be obscured, and therefore less likely to be detected. The analysis with high noise levels, greater than −3100 dB (**Section** 3.2), will be carried out in the future.

TABLE 16. Comparison of proposed localization results with existing techniques using Gunshot dataset (locally recorded dataset have been used by the researchers, while we used the publically Available Acoustic dataset); NM stands for not mentioned.

Ref.	No. of Mic	Distance	Geometry	Noise	A_{loc} (%)
[9]	NM	NM	NM	×	97.00
[8]	6	Evenly distributed (30°)	Circular	√	93.00
[3]	4	30 cm	T-shaped	√	93.00
[7]	1	NM	NM	×	96.00
[6]	6	Evenly distributed (30°)	Circular	√	97.00
Proposed	4	70 cm	Omni-directional	√	99.95

TABLE 17. Comparison of proposed classification results with existing techniques using original dataset.

Ref.	**Method**	**F-score**	**A_{cls} (%)**
[31, 89]	PNCC+SVM	0.46	69.02
[33, 34]	DWT+ELM	0.44	77.62
[28, 29]	GFCC+ELM	0.60	85.08
[35]	(DWT-MFCC)+SVM	0.72	90.21
[36]	(MFCC-GFCC)+SVM	0.72	90.21
[37]	(MFCC-PNCC)+SVM	0.73	90.91
Proposed	1D-Mild CNN+MFCC	0.79	88.42
Proposed	**MFCC+ELM**	**0.83**	**93.01**

The effect of echo in gunshots can only be analyzed in known environmental conditions. In the future, the effect of echo will be investigated in real-time scenarios. Furthermore, the gun position in different directions either raised higher or lower to the microphones, will also be figured out in future studies. The robustness of the classification model, although has been found to outclass against wind noise conditions, however, other environmental noises like loud door slams, and handclaps need to be investigated as the detection accuracy of false alarms.

IV. CONCLUSION

In this work, a framework is proposed for the localization and classification of gunshots in normal and strong windy environments for a surveillance system. The sensitivity analysis of normal and strong wind noise impregnated acoustic signals has been conducted for gunshot events to localize the acoustic event source and devise an ML model for the classification of the imbalanced dataset. The filtration of the simulated and noise impregnated acoustic signals has been carried out using conventional as well as adaptive filters. It has been observed through experiments that the reconstructed signal quality using median filter, among conventional filters' class, is relatively more robust against specified wind noise. However, the acoustic signals based on adaptive filters quality are much improved as compared with the conventional filters. In the adaptive filters class, the performance of wavelet transforms-based denoising filters has shown relatively promising results. Further, localization of the acoustic source using Hadamard product approach in combination with wavelet de-noising has been used first time for the localization of gunshots in windy conditions. We have found that the proposed frequency domain approach relatively outperforms with a localization accuracy of 99.95%. It has been found that the classification using ELM, not known before for gunshots dataset to the best of our knowledge, has been found robust with the classification performance in terms of accuracy for original gunshots, normal- and strong-wind noise impregnated events as (93.01%, 91.61%, and 88.11%) respectively.

REFERENCES

[1] V. Blondeau-Patissier, M. Vanotti, E. Quivet, and D. Buiron, "Surface acoustic wave sensors for fine particle detection air quality monitoring," in *Proc. 7th Int. Conf. Sensor Device Technol. Appl.* Nice, France, 2016, pp. 15–16.

[2] A. P. Sarvazyan, M. W. Urban, and J. F. Greenleaf, "Acoustic waves in medical imaging and diagnostics," *Ultrasound Med. Biol.*, vol. 39, no. 7, pp. 1133–1146, Jul. 2013.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, Sep. 2007, pp. 21–26.

[4] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Appl. Acoust.*, vol. 117, pp. 207–218, Feb. 2017.

[5] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, Sep. 1994.

[6] S. Astapov, J. Ehala, J. Berdnikova, and J.-S. Preden, "Gunshot acoustic component localization with distributed circular microphone arrays," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 1186–1190.

[7] N. B. Gaikwad, H. Ugale, A. Keskar, and N. C. Shivaprakash, "The internet-of-battlefield-things (IoBT)-based enemy localization using soldiers location and gunshot direction," *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11725–11734, Dec. 2020.

[8] S. Astapov, J. Berdnikova, J. Ehala, J. Kaugerand, and J.-S. Preden, "Gunshot acoustic event identification and shooter localization in a WSN of asynchronous multichannel acoustic ground sensors," *Multidimensional Syst. Signal Process.*, vol. 29, no. 2, pp. 563–595, Apr. 2018.

[9] N. Pathrose, K. R. Nair, R. Murali, K. R. Rajesh, N. Mathew, and S. Vishnu, "Analysis of acoustic signatures of small firearms for gun shot localization," in *Proc. IEEE Annu. India Conf. (INDICON)*, Dec. 2016, pp. 1–5.

[10] Y. Anzai, *Pattern Recognition and Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2012.

[11] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[12] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[13] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: A survey," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 2, pp. 107–122, Jun. 2011.

[14] S. D. Correia, S. Tomic, and M. Beko, "A feed-forward neural network approach for energy-based acoustic source localization," *J. Sensor Actuator Netw.*, vol. 10, no. 2, p. 29, Apr. 2021.

[15] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, Oct. 2018.

[16] S. T. H. Shah, S. A. Qureshi, A. U. Rehman, S. A. H. Shah, A. Amjad, A. A. Mir, A. Alqahtani, D. A. Bradley, M. U. Khandaker, M. R. I. Faruque, and M. Rafique, "A novel hybrid learning system using modified breaking ties algorithm and multinomial logistic regression for classification and segmentation of hyperspectral images," *Appl. Sci.*, vol. 11, no. 16, p. 7614, Aug. 2021.

[17] U. Sharma, S. Maheshkar, and A. N. Mishra, "Study of robust feature extraction techniques for speech recognition system," in *Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manage. (ABLAZE)*, Feb. 2015, pp. 654–658.

[18] L. Cao, K. Chua, W. Chong, H. Lee, and Q. Gu, "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, nos. 1–2, pp. 321–336, 2003.

[19] I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems With Applications in ehealth, HCI, Information Retrieval and Pervasive Technologie*, vol. 160. Amsterdam, The Netherlands: Ios Press, 2007.

[20] D. M. Abdullah and A. M. Abdulazeez, "Machine learning applications based on SVM classification a review," *Qubahan Academic J.*, vol. 1, no. 2, pp. 81–90, Apr. 2021.

[21] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Mater. Today, Proc.*, vol. 7, p. 361, Aug. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785321052202?via%3Dihub, doi: 10.1016/j.matpr.2021.07.361.

[22] S. S. Ye and O. H. M. Padilla, "Non-parametric quantile regression via the K-NN fused lasso," *J. Mach. Learn. Res.*, vol. 22, no. 111, pp. 1–38, 2021.

[23] Y. Wei, Y. Yang, M. Xu, and W. Huang, "Intelligent fault diagnosis of planetary gearbox based on refined composite hierarchical fuzzy entropy and random forest," *ISA Trans.*, vol. 109, pp. 340–351, Mar. 2021.

[24] X. Liu, D. Pei, G. Lodewijks, Z. Zhao, and J. Mei, "Acoustic signal based fault detection on belt conveyor idlers using machine learning," *Adv. Powder Technol.*, vol. 31, no. 7, pp. 2689–2698, Jul. 2020.

[25] S. Dabetwar, S. Ekwaro-Osire, and J. P. Dias, "Damage classification of composites based on analysis of Lamb wave signals using machine learning," *ASCE-ASME J. Risk Uncertainty Eng. Syst., B, Mech. Eng.*, vol. 7, no. 1, Mar. 2021, Art. no. 011002.

[26] Q.-V. Pham, N. T. Nguyen, T. Huynh-The, L. Bao Le, K. Lee, and W.-J. Hwang, "Intelligent radio signal processing: A survey," *IEEE Access*, vol. 9, pp. 83818–83850, 2021.

[27] B. Rozemberczki, P. Scherer, Y. He, G. Panagopoulos, A. Riedel, M. Astefanoaei, O. Kiss, F. Beres, G. López, N. Collignon, and R. Sarkar, "PyTorch geometric temporal: Spatiotemporal signal processing with neural machine learning models," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 4564–4573.

[28] X. Shi, H. Yang, and P. Zhou, "Robust speaker recognition based on improved GFCC," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2016, pp. 1927–1931.

[29] Z. Lian, K. Xu, J. Wan, and G. Li, "Underwater acoustic target classification based on modified GFCC features," in *Proc. IEEE 2nd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2017, pp. 258–262.

[30] A. Maesa, "Text independent automatic speaker recognition system using mel-frequency cepstrum coefficient and Gaussian mixture models," *J. Inf. Secur.*, vol. 3, no. 4, p. 335, 2012.

[31] R. Thiruvengatanadhan, "Speech/music classification using power normalized cepstral coefficients and K-means," *Int. J. Adv. Comput. Sci. Technol.*, vol. 6, no. 1, pp. 13–17, 2016.

[32] M. T. Al-Kaltakchi, "Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database," *Indonesian J. Elect. Eng. Comput. Sci.*, vol. 18, no. 2, pp. 586–598, 2020.

[33] S. Desai and M. A. H. Morcos, "Classifying launch/impact events of mortar and artillery rounds utilizing DWT-derived features and feedforward neural networks," *Proc. SPIE*, vol. 6247, pp. 227–238, Apr. 2006.

[34] Z.-L. Zhou, R.-S. Cheng, L.-J. Chen, J. Zhou, and X. Cai, "An improved joint method for onset picking of acoustic emission signals with noise," *J. Central South Univ.*, vol. 26, no. 10, pp. 2878–2890, Oct. 2019.

[35] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, and G. R. Naik, "Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions," *IEEE Access*, vol. 5, pp. 15400–15413, 2017.

[36] E. B. Tazi, "A robust speaker identification system based on the combination of GFCC and MFCC methods," in *Proc. 5th Int. Conf. Multimedia Comput. Syst. (ICMCS)*, Sep. 2016, pp. 54–58.

[37] K. P. Bharath, "ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 28859–28883, Oct. 2020.

[38] C. M. Nelke, *Wind Noise Reduction: Signal Processing Concepts*. Mainz. Germany: Wissenschaftsverlag Mainz, 2016.

[39] R. Schafer, "What is a Savitzky–Golay filter? [lecture notes]," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 111–117, Jul. 2011.

[40] A. Al-Odienat and A. Al-Mbaideen, "Optimal length determination of the moving average filter for power system applications," *Int. J. Innov. Comput., Inf. Control*, vol. 11, no. 2, pp. 691–705, Apr. 2015. [Online]. Available: http://www.ijicic.org/contents.htm

[41] S. C. Douglas, "Introduction to adaptive filters," in *Digital Signal Processing Handbook*, V. K. Madisetti and D. B. Williams, Eds. Boca Raton, FL, USA: CRC Press, 1999.

[42] D. Bhoyar and P. Singh, "ECG noise removal using adaptive filtering," in *Proc. Int. Conf. Ind. Automat. Comput.*, 2014, pp. 21–24.

[43] J. Dhiman and S. K. A. Gulia, "Comparison between adaptive filter algorithms (LMS, NLMS and RLS)," *Int. J. Sci., Eng. Technol. Res.*, vol. 2, no. 5, pp. 1100–1103, 2013.

[44] S. Khan and M. T. A. Majeed, "Comparison of LMS, RLS and notch based adaptive algorithms for noise cancellation of a typical industrial workroom," in *Proc. 8th Int. Multitopic Conf.*, Dec. 2004, pp. 169–173.

[45] R. A. Goubran, R. Herbert, and H. M. Hafez, "Acoustic noise suppression using regressive adaptive filtering," in *Proc. 40th IEEE Conf. Veh. Technol.*, May 1990, pp. 48–53.

[46] C. Breining, "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[47] R. K. Thenua and S. Agarwal, "Simulation and performance analysis of adaptive filter in noise cancellation," International Journal of Engineering Science and Technology, 2010. vol. 2, no. 9, pp. 4373–4378.

[48] D. R. Raichel, *The Science and Applications of Acoustics*. Cham, Switzerland: Springer, 2006.

[49] (Oct. 2021). *Still North Media*. Accessed: Feb. 2021. [Online]. Available: https://www.stillnorthmedia.com/libraries

[50] B. Berdugo, M. A. Doron, J. Rosenhouse, and H. Azhari, "On direction finding of an emitting source from time delays," *J. Acoust. Soc. Amer.*, vol. 105, no. 6, pp. 3355–3363, Jun. 1999.

[51] H.-W. Lee, J.-W. Lee, W.-G. Jung, and G.-K. Lee, "The periodic moving average filter for removing motion artifacts from PPG signals," *Int. J. Control Autom. Syst.*, vol. 5, no. 6, pp. 701–706, Dec. 2007.

[52] W. Mikhael, F. Wu, L. Kazovsky, G. Kang, and L. Fransen, "Adaptive filters with individual adaptation of parameters," *IEEE Trans. Circuits Syst.*, vol. CS-33, no. 7, pp. 677–686, Jul. 1986.

[53] S. Guan, Q. Cheng, Y. Zhao, and B. Biswal, "Diffusion adaptive filtering algorithm based on the fair cost function," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Dec. 2021.

[54] Y. Bai, X. Wang, X. Jin, T. Su, J. Kong, and B. Zhang, "Adaptive filtering for MEMS gyroscope with dynamic noise model," *ISA Trans.*, vol. 101, pp. 430–441, Jun. 2020.

[55] A. Parshin and Y. Parshin, "Adaptive filtering of non-Gaussian flicker noise," in *Proc. 9th Medit. Conf. Embedded Comput. (MECO)*, Jun. 2020, pp. 1–5.

[56] G. Xing and Y. Zhang, "Analysis and comparison of RLS adaptive filter in signal de-noising," in *Proc. Int. Conf. Electr. Control Eng.*, Sep. 2011, pp. 5754–5758.

[57] C. R. Cornish, C. S. Bretherton, and D. B. Percival, "Maximal overlap wavelet statistical analysis with application to atmospheric turbulence," *Boundary-Layer Meteorol.*, vol. 119, no. 2, pp. 339–374, May 2006.

[58] C. Taswell, "The what, how, and why of wavelet shrinkage denoising," *Comput. Sci. Eng.*, vol. 2, no. 3, pp. 12–19, May 2000.

[59] T. C. Tran and M. N. H. H. B. Nguyen, "A modified localization technique for pinpointing a gunshot event using acoustic signals," in *Proc. Int. Conf. Ind. Netw. Intell. Syst.* Cham, Switzerland: Springer, 2020, pp. 138–149.

[60] T. Padois, O. Doutres, F. Sgard, and A. Berry, "Time domain localization technique with sparsity constraint for imaging acoustic sources," *Mech. Syst. Signal Process.*, vol. 94, pp. 85–93, Sep. 2017.

[61] B. Bin, L. Guo-Chun, L. Tao, L. Yu-Cheng, and W. Yu, "Joint for time of arrival and direction of arrival estimation algorithm based on the subspace of extended Hadamard product," *Acta Phys. Sinica*, vol. 64, no. 7, 2015, Art. no. 078403.

[62] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, ''Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments,'' *IEEE Access*, vol. 8, pp. 7373–7382, 2020.

[63] N. S. Faeza, J. Chunkath, N. Pathrose, and R. Kr, ''Identification of shockwave and muzzle blast in a gunshot signal using frequency analysis techniques,'' in *Proc. Int. Conf. Power, Instrum., Control Comput. (PICC)*, Dec. 2020, pp. 1–4.

[64] M. Cobos, F. Antonacci, L. Comanducci, and A. Sarti, ''Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1270–1281, 2020.

[65] L. Fostick and N. Fink, ''Situational awareness: The effect of stimulus type and hearing protection on sound localization,'' *Sensors*, vol. 21, no. 21, p. 7044, Oct. 2021.

[66] S. Dogan, ''A new fractal H-tree pattern based gun model identification method using gunshot audios,'' *Appl. Acoust.*, vol. 177, Jun. 2021, Art. no. 107916.

[67] R. B. Singh, H. Zhuang, and J. K. Pawani, ''Data collection, modeling, and classification for gunshot and gunshot-like audio events: A case study,'' *Sensors*, vol. 21, no. 21, p. 7320, Nov. 2021.

[68] L. Gerosa, ''Scream and gunshot detection in noisy environments,'' in *Proc. 15th Eur. Signal Process. Conf.*, Sep. 2007, pp. 1216–1220.

[69] S. Guo, Y. Liu, R. Chen, X. Sun, and X. Wang, ''Improved SMOTE algorithm to deal with imbalanced activity classes in smart Homes,'' *Neural Process. Lett.*, vol. 50, no. 2, pp. 1503–1526, Oct. 2019.

[70] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, ''Real-time acoustic based depression detection using machine learning techniques,'' in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (IC-ETITE)*, Feb. 2020, pp. 1–6.

[71] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, ''Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients,'' in *Proc. IEEE Conf. Evolving Adapt. Intell. Syst. (EAIS)*, May 2020, pp. 1–5.

[72] S. Raj and P. S. P. Gupta, ''Multilayered convolutional neural network-based auto-CODEC for audio signal denoising using mel-frequency cepstral coefficients,'' *Neural Comput. Appl.*, vol. 33, no. 16, pp. 10199–10209, 2021.

[73] A. I. Siam, A. A. Elazm, N. A. El-Bahnasawy, G. M. El Banby, and F. E. A. El-Samie, ''PPG-based human identification using mel-frequency cepstral coefficients and neural networks,'' *Multimedia Tools Appl.*, vol. 80, no. 17, pp. 26001–26019, Apr. 2021.

[74] D. Namrata, ''Feature extraction methods LPC, PLP and MFCC in speech recognition,'' *Int. J. Advance Res. Eng. Technol.*, vol. 1, no. 6, pp. 1–4, 2013.

[75] M. S. B. A. Ghaffar, U. S. Khan, J. Iqbal, N. Rashid, A. Hamza, W. S. Qureshi, M. I. Tiwana, and U. Izhar, ''Improving classification performance of four class FNIRS-BCI using mel frequency cepstral coefficients (MFCC),'' *Infr. Phys. Technol.*, vol. 112, Jan. 2021, Art. no. 103589.

[76] Ö. F. Ertuğrul and Y. Kaya, ''A detailed analysis on extreme learning machine and novel approaches based on ELM,'' *Amer. J. Comput. Sci. Eng.*, vol. 1, no. 5, pp. 43–50, 2014.

[77] A. Indumathi and E. Chandra, ''An efficient speaker recognition system by employing BWT and ELM,'' *BVICA M's Int. J. Inf. Technol.*, vol. 8, no. 2, p. 983, 2016.

[78] I. Rish, ''An empirical study of the naive Bayes classifier,'' in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, pp. 41–46.

[79] U. I. Awan, U. H. Rajput, G. Syed, R. Iqbal, I. Sabat, and M. Mansoor, ''Effective classification of EEG signals using K-nearest neighbor algorithm,'' in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2016, pp. 120–124.

[80] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2020.

[81] G. Biau, ''Analysis of a random forests model,'' *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, Apr. 2014.

[82] D. Poobathy and R. M. Chezian, ''Edge detection operators: Peak signal to noise ratio based comparison,'' *Int. J. Image, Graph. Signal Process.*, vol. 6, no. 10, pp. 55–61, Sep. 2014.

[83] A. Rahaman, C. H. Park, and B. Kim, ''Design and characterization of a MEMS piezoelectric acoustic sensor with the enhanced signal-to-noise ratio,'' *Sens. Actuators A, Phys.*, vol. 311, Aug. 2020, Art. no. 112087.

[84] J. Bradley, R. D. Reich, and S. Norcross, ''On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility,'' *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1820–1828, 1999.

[85] X. R. Li and Z. Zhao, ''Relative error measures for evaluation of estimation algorithms,'' in *Proc. 7th Int. Conf. Inf. Fusion*, Jul. 2005, p. 8.

[86] V. Labatut and H. Cherifi, ''Accuracy measures for the comparison of classifiers,'' 2012, *arXiv:1207.3790*.

[87] T. Pál and D. T. Várkonyi, ''Comparison of dimensionality reduction techniques on audio signals,'' in *Proc. ITAT*, 2020, pp. 161–168.

[88] J. Singh and R. Joshi, ''Background sound classification in speech audio segments,'' in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2019, pp. 1–6.

[89] E. Ambikairajah, ''PNCC-Ivector-SRC based speaker verification,'' in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2012, pp. 1–7.