JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 02
2025

*Research Article*
https://doi.org/10.56979/902/2025

# Adaptive Boosted Support Vector Machine-random Forest for Environmental Sound Classification

**Faiz Ul Hasnain[1*], Visha Iqbal[2], Tayyaba Javed[3], and Muhammad Yasir[2]**

[1]Department of Computer Science, University of Education Lahore, Vehari Campus, Vehari, Pakistan.
[2]Department of Computer Science, University of Management and Technology Lahore, Pakistan.
[3]Department of Computer Science, Barani Institute of Information Technology, Rawalpindi 46604, Pakistan.
[*]Corresponding Author: Faiz Ul Hasnain. Email: faizhaans4@gmail.com

**Received:** July 10, 2025 **Accepted:** August 23, 2025

**Abstract:** Environmental sound classification (ESC) is a method to differentiate the audio related to the various environmental sounds. Environmental sounds have a more complex time-frequency structure compared to structured sounds like music and speech. To extract the frequency and time-based features from audio more accurately and effectively, a novel fusion of several features including MFCCs, Mel-spectrogram, spectral skewness, spectral kurtosis and normalized pitch frequency will be evaluated in this study to provide a comprehensive representation of environmental sounds. The fusion will capture various aspects of the input audio data, such as spectral characteristics, statistical properties, and frequency-related information. By using multimodal information fusion, the algorithm will enhance the discriminative power of the model to distinguish between different sounds more effectively. Moreover, the integration of a variety of machine learning models will enhance the robustness and generalization ability of the model. The combination of several machine learning models will reduce the training time and enhance the classification rate of environmental audio under limited computational resources. Furthermore, this thesis will employ three data augmentation methods, namely, time stretch, pitch tuning, and white noise to minimize the probability of overfitting problems due to the limited audios in each class of dataset. This research will evaluate the ensemble model classification accuracy against baseline SVM, RF classifiers, and other state-of-the-art approaches. In UrbanSound8K, ESC-50, and ESC-10 datasets, the highest achieved accuracies using AdaBoost SVM-RF classifiers were ( 94%), (85%), and ( 95%) respectively. The experimental findings demonstrate that the suggested approach achieves superior performance for ESC tasks.

**Keywords:** ESC; Environmental Sound; UrbanSound8K; ESC10; ESC50; AdaBoost; Feature Fusion

## 1. Introduction

Signal processing extracts meaningful patterns like frequency and amplitude from raw audio signals. Deep learning models then learn complex features directly from this processed data. Together, they enhance the accuracy and efficiency of audio analysis. This powerful combination is widely used in tasks such as human activity recognition [1], speaker identification [2, 3], speech emotion recognition [4] and environmental sound classification [5]. Environmental sound refers to the natural or artificial sounds that are present in a particular environment. These sounds are from two sources natural and human-made sound. Animal sounds, birdsong, insect sounds, and elements of nature such as wind, water flowing, waves crashing, and thunder, and rain, make up natural sounds. Noises made by humans include traffic noise, construction noise, and industrial noise. Natural and man-made sounds combine to produce environmental sounds. The sound of distant traffic might be heard along with birds singing in a city park. There are many types of environmental sounds, including natural and artificial sounds. Their effects are

felt by wildlife, humans, and society as a whole. Noise pollution can be reduced through mitigation methods and regulations.

Environmental Sound Classification (ESC) uses advanced algorithms and machine learning to recognize and classify a wide variety of nonspeech, and non-musical sounds as shown in Figure 1. This field has been crucial to the development of intelligent technology, urban planning, and animal monitoring [6]. It classifies sounds from our environment as part of audio signal processing. During the first step, data was collected, including background noises like traffic, equipment, rain, and animal sounds as well as man-made sounds. After preprocessing, this audio material would be cleaned and normalized. Depending on the situation, it may be necessary to reduce background noise, increase loudness, or break up long audio samples [7]. While analyzing audio, important features such as spectral contrast, zero-crossing rates, and Mel-frequency cepstral coefficients (MFCCs) are extracted [8]. Sound classification requires these characteristics. That was equally significant in selecting the right model. Support Vector Machines (SVMs) and Random Forests are examples of advanced deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [9]. Following the preprocessing of the data, the chosen model was trained and features associated with different sound classes were extracted. Testing and validation evaluate the model's accuracy and generalizability by comparing it with newly collected, and untested data. Filtering and class combination may be required to improve the findings, post-processing techniques. The system aims for user engagement and will require a simple user interface [10]. Retrained frequently to stay effective the system needs to be updated. It must also have the ability to adjust to new sounds and changing surroundings. The precision, reliability, and usefulness of the ambient sound categorization system were all dependent on each of these components [11].

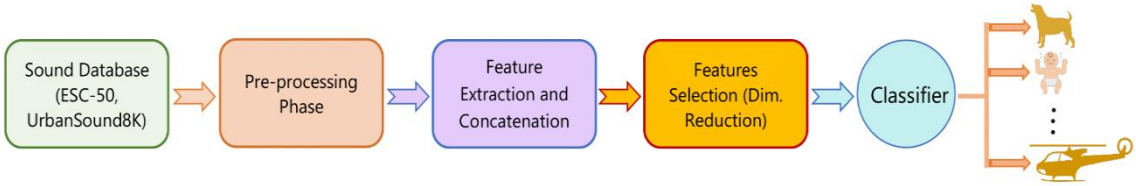



**Figure 1.** Classical Process of ESC.

Classification of noises outside typically involves optimal exploitation of acoustic features and improved classifier implementation. Using a hamming window, the audio signal typically split into frames for extract the audio features. Testing and training are based on characteristics taken from each frame [12]. The richness and diversity of audio sources pose unique problems for environmental noise categorization. Compiling large databases can be challenging because noises originate from a variety of sources, including weather patterns, animals, and human activities. The main problem was background noise, which often drowns out fascinating sounds [13]. There can be difficulty distinguishing particular sounds from background noise because ambient noises are often very varied and overlapped. Additionally, there aren't many large, well-labeled datasets available for supervised machine learning [14]. A variety of temporal characteristics are observed in the noises, with some being brief and transient and others being prolonged [15]. It was also difficult to interpret signals using traditional methods because the characteristics of these sounds change over time. Using sophisticated signal processing and robust machine learning techniques is required to resolve problems related to ambient sound categorization. A range of standardized datasets was essential for ensuring recording quality [16]. Employing crowdsourcing for data labeling, leveraging computing power for real-time processing, and fostering interdisciplinary collaboration can enhance effectiveness. Additionally, addressing privacy concerns with strict guidelines and sound anonymization is essential, particularly in urban settings, to ensure ethical practices in sound collection and analysis [17].

The main challenge in environmental sound classification systems is addressing the complexity and variability of environmental sounds. These sounds exhibit substantial diversity in context, intensity, and duration, posing challenges for constructing accurate and reliable classification systems. Additionally, the selection and placement of recording devices significantly impact the effectiveness of these systems. For example, strategically placed recording devices can yield precise data whereas less accurate devices might offer recordings at the cost of data accuracy. There is growing interest in developing environmental sound classification systems that can operate in real-world settings characterized by unstructured and unpredictable sounds. One strategy to tackle this issue is integrating multimodal sounds and extracting

the combination of several features to capture various aspects of the input audio data, such as spectral characteristics, statistical properties, and frequency-related information. Secondly, computational efficiency is crucial for real-time applications, necessitating the development of algorithms that can process data quickly without reducing accuracy.

This study presents a new approach to ESC by proposing the Adaptive boosted support vector machine-Random Forest model, set a new standard for computational acoustics and machine learning. This research points out the input and output of this model, which are to create an extensive and multipurpose collection of data on environmental sounds necessary for the models training and testing. The innovation used in the research is the applying and optimizing of the Adaptive Boosted SVM-RF to manage difficult factors such as dissimilar sound and overlapping noise as part of the development in audio data analysis. Sophisticated audio data preprocessing methods are derived to process audio information effectively to improve its responsiveness to the model. For instance, the elimination of background noise and feature extraction. The goal of the work is to exceed the existing algorithms in their true accuracy and contextual analysis. Hence applying the boosting methodology in the SVM-RF to improve the outcome. For the same manner, objective measures and evaluation parameters are set solid in both extensive and sound relevant metrics for further studies and enhancements.

The rest of this study is divided as follows. Section 2 provides the comprehensive literature review that includes all the mechanisms and methods related to the research. In section 3 a detailed methodology of the research is elaborated. In sction 4 the experimental results are discussed and evaluated. Finally, section 5 reports the conclusion and future work.

## 2. Literature Review

Environmental sound classification includes artificial sounds (office noises, urban noise) and natural sounds (wildlife, weather). Each presents unique challenges, such as diverse, unstructured data and background noise. Effective feature extraction, like Mel spectrograms and MFCCs, is crucial. Discussion in detail about Sound sources, ESC techniques, Challenges and considerations and Research direction are given below as shown in Figure 2.

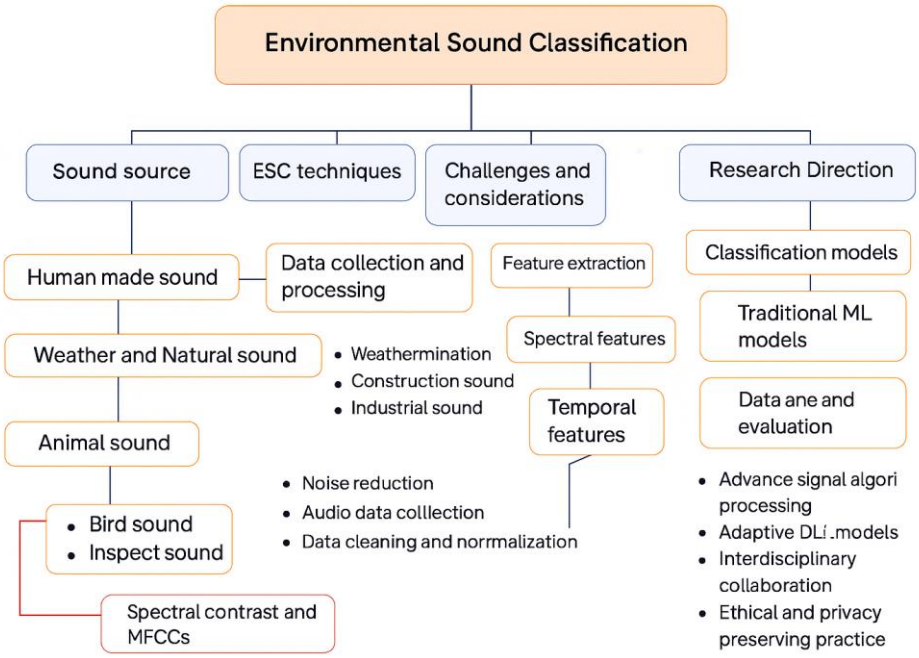



**Figure 2.** Review Taxnomy of ESC.

### 2.1. Sound Sources of ESC

The study [18] demonstrates how a BERT-based Transformer, appropriate for edge devices, effectively identifies office noises with 99.85% fewer parameters than CNNs. Small datasets, and more recently noise, remain the problem. For en hancing the effectiveness and efficiency of edge computing applications in AAC, there are several research directions to be explored in the future These include deploying models on-

device and fine-tuning, and also designing models for differ ent sound sources. The study [19] propose a CNN-based approach that employs FDM-converted audio data in the Mel-spectrogram and Scalogram domains for urban sound classification and ambulance siren detection. It outperforms other pretrained models including GoogleNet, AlexNet and VGG-16 with a score of 89. For classifying urban noises, the network has an accuracy of 66% and for sepa rating hearing aids' noise from other types of noise, it has an accuracy of 99%. 35% accuracy in recognizing siren similar to the previous experiment, but here, the results clearly show the effectiveness of data augmentation and pre-processing data methods. While the study clearly discusses the strength of the model in terms of flexibility, accuracy and capability to be tested using the existing Korean data, the same work could also denote the factors that are actually associated with its complexity, data dependence, and generalization. It may have potential in fields such as medical monitoring and traffic management especially if more tweaks, which would provide extra performance and versatility, can be made to the technique later on. This study [20] address the challenges faced by the users of CI in identifying noises in their environment by employing CNN resources for sound identification models. Such controls for normal hearing (NH) and conditional inhibition (CI) are created for both natural and human-made sounds. This study shows that the current generation of CI technology is limited in the frequency and temporal accuracy, causing users to struggle in perceiving distinctions of various tones. The given ability to recognize the sounds of the weather, wind and water provides not only an enhancement to the quality of the thought process but also facilitates a benefit towards general well-being and the environment. To enhance the quality of the user experience and later suggest a definite boost in the quality of life of CI users The study proposed [21] Results Fine, let me tell you that the ability to recognize sounds in the environment, called Environmental Sound Recognition, or ESR, is crucial both in monitoring the urban environment and in the protection of animals. The use of ESR for object detection is improved by deep learning particularly by CNN. TPUs and FPGAs all operate within computational constraints, which are an inherent aspect of embedded devices.

2.2. Deep Learning techniques for ESC

This study [22] investigates the susceptibility of deep learning models for envi ronmental sound classification to adversarial attacks. It introduces the Adv-ESC dataset, crafted from ESC-10 and DCASE 2019 Task-1A Challenge datasets, incor porating three prevalent attack techniques: FGSM, BIM, and PGD. Adversarial perturbations are applied to spectrogram features to simulate real-world scenarios. Deep models trained on Adv-ESC exhibit significant accuracy degradation when exposed to adversarial examples. This dataset enables benchmarking of model ro bustness against attacks and highlights the necessity of developing resilient sound classification systems. The findings emphasize the importance of addressing adver sarial vulnerabilities in environmental sound classification. This author [23] pro posed that audio data collection methods, ranging from traditional microphones to wearable devices, are fundamental in machine learning and signal processing. The quality of the hardware, the surrounding environment, and the synchronization strategies all affect accuracy. Difficulties include privacy problems, short battery life, and background noise. Data authenticity has to be maintained despite im provements. To overcome constraints and improve data quality for a range of applications, interdisciplinary cooperation is required. This study [24] presents an alternative method to improve the accuracy of ambient sound categorization. With the use of a time-frequency attention mechanism and the integration of multi feature characteristics, such as phase spectrograms, the technique is able to record complex sound patterns with little interference from background noise. Exten sive tests on the ESC-10, ESC-50, and UrbanSound8K datasets confirm improved performance above current techniques. With creative feature representation and noise reduction strategies, the suggested model shows promise for improving am bient sound identification systems and is resilient in noisy conditions. The author [25] demonstrates the need of data normalization and cleaning for reliable ambient sound categorization. Normalization normalizes features; cleaning corrects errors and removes outliers. When combined with neural networks such as CNNs and RNNs, these methods improve the accuracy of the models. Dealing with data un predictability and standardization present challenges. Subsequent investigations attempt to establish standardized approaches and investigate unsupervised learn ing with practical implications. This study proposed [26] that the process of turn ing unprocessed audio into machine-understandable representations in environ mental sound categorization (ESC) requires the extraction of features. Capturing complex patterns is

facilitated by transfer learning, which makes use of pre-trained convolutional neural networks (CNNs) such as Inception, VGG, or ResNet. Spec trograms are useful feature spaces, particularly log Mel-spectrograms. CNNs are optimized by programs like Adam, Adamax, and RMSprop for ESC tasks. The UrbanSound8K dataset, which offers a variety of urban noises, has been frequently utilized in evaluation. In order to improve feature extraction, future studies may examine attention processes. All things considered, using pre-trained CNNs and f ine-tuning hyperparameters provide encouraging outcomes, signifying a substan tial advancement in ESC. Convolutional Neural Networks (CNNs) have issues in Environmental Sound Classification (ESC) because of growing model complexity.

The ESC system has quite a lot of aspects in terms of performance improve ments comparing with the previous methods based on the handmade features and ML correlates and resulted in deep learning. The proposed study [27] pro vides a new solution for increasing the accuracy and robustness of the classifiers that employs attention procedures, connectionist CNN models, and spectrogram based features. The method needs several pre-trained models and improves data augmentation to capture detailed features of the audio signal. The files of the Urbansound8k dataset experiment show the result which is higher than baseline models' results. It therefore has pros and cons as follows: The advantage of the approach is that it is purely computational, and is relatively simple to implement by running the function and adjusting parameters. The proposed study [28] us ing CNNs and RNNs, sound recognition proves to outdo traditional approaches in efficiencies since they find attributes from the sound frequency. Transformers and attention mechanisms are used to create adaptive models that perform even better due to focus on relevant segments of audio. Even while these models are able to achieve above 80% accuracy in tasks such as ambient sound categoriza tion, they have drawbacks such as high computing needs and the demand for large annotated datasets. Challenges in generalizing across varied acoustic envi ronments persist, necessitating efficient model development and robust training methodologies for practical, real-world applications.

It is clear that improving the data augmentation and developing deeper CNN mod els can both improve the ESC prediction performance. On the other hand, the training model has been limited to roughly fifteen convolutional layers due to the typical moving dataset sizes. However, excellent results on ESC tasks have been shown by the transfer learning technique employing a pre-trained model on Ima geNet, which makes it possible to obtain a higher outcome to train deeper models that can easily. As audio classifiers are trained using spec trograms, which show tight relationships between local locations, using pretrained models on ImageNet improves feature extraction and predictive accuracy.

## 3. Materials and Methods

The proposed Adaptive boosted support vector Machine-random Forest for ESC will be collected in five steps as shown in Figure 3. Proposed the complexities of the databases and will use in our studies, feature extraction strategies, data augmentation approaches, Adaptive boosted Support vector-machine random forest, and metrics used to assess performance in the proposed methodology. Gathering relevant sound information is the first step in the current sound classification process. Evaluated machine learning model on three datasets: ESC-10, ESC-50, and UrbanSound8K for ESC. The proposed methodology will use Data Augmentation (DA) techniques like pitch shift, Gaussian noise, and time stretch to expand the training dataset and reduce overfitting. There is a large amount of data necessary to train a ML model for ESC. This research also describes the features of MFCCs, Chromagram, Mel-spectrogram, delta-delta MFCCs, Tonnetz, Spectral Contrast, and delta-MFCCs representation that will be used to train the suggested model. This research provides a complex support vector Machine-random Forest model. Lastly, this research clarifies the assessment metrics used to evaluate our suggested model's performance.

### 3.1. Dataset

This research uses three data sets that anyone can access to teach the computer program and check how well our new method works. These datasets are called UrbanSound8K, ESC-10, and ESC-50. The UrbanSound8K database has 8,732 small audio files (lasting up to 4 seconds) that capture sounds from urban areas. The information in the datasets is divided into 10 groups: engine, idling, siren, jackhammer, street music, car horn, air conditioner, children playing, drilling, and dog bark. The class recordings are split into 10 groups. In total, there are 400 sound files in the ESC-10 collection, and each file is about 5

seconds long on average. These sound recordings include 10 different types of sounds, like rain, a dog barking, a baby crying, sea waves, a clock ticking, a helicopter, a person, a sneeze, a rooster, a chainsaw, and fire crackling. In total, they last for 33 minutes. This data is split into five parts. Each part has 80 audio files with different types of sounds randomly mixed in. 2000 short audio files are stored in the ESC-50 database. They are sorted into 50 groups, and these groups are divided into 5 main sets. The sets include things like water sounds and natural scenes, animals, human sounds that aren't speech, city and outdoor noises, and sounds from inside homes. This collection of information will split into five parts, and each part has 400 sound files. All the sound recordings in ESC-50 and ESC-10 were made with a sampling frequency of 44.1 kilohertz, and each recording will be 5 seconds long.

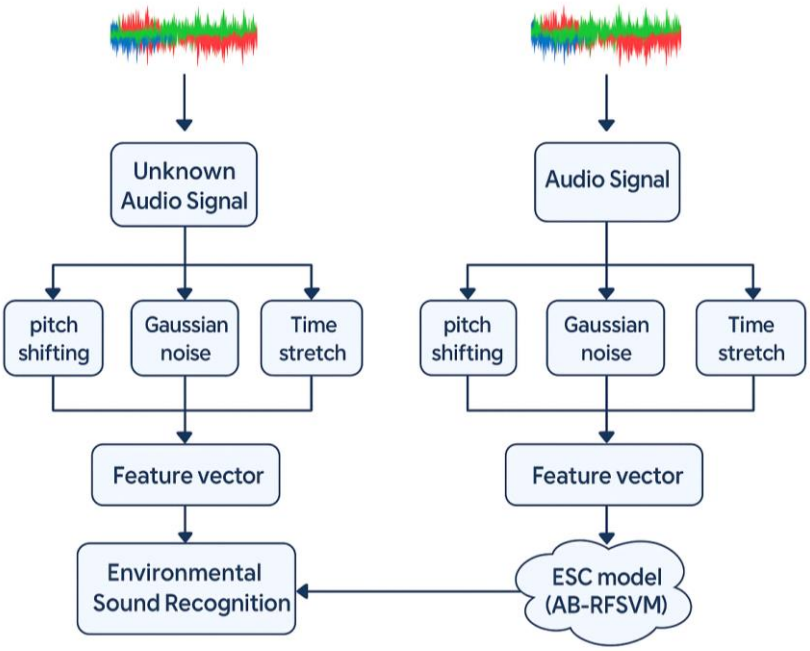



**Figure 3.** Proposed Research Methodology

3.2. Data Enhancement

One the way people often prevent machine learning models from getting too specialized and not working well in different situations when classifying environmental sounds by using something called data enhancement. Data enhancement means making the training data bigger by doing something like adding noise, changing pitch, and stretching time. Data enhancement makes the deep learning model stronger by improving its ability to understand different situations. It helps the model become better at handling a variety of data, making it more accurate overall. Additionally, it evens out the distribution of data and reduces variations in the data. First, we made the training data bigger by adding a type of random noise called Gaussian noise. When you're adding more training data with Gaussian noise, an important thing to think about is how much noise you're putting in, which is called the (δ) value. It's important to pick the right value for (δ). If the number for 'a' is too big, it can be hard to make the model work well. On the other hand, if 'a' is too small, the model might not perform well either. Next, we create new sounds by changing the pitch and shifting the tone up or down in a certain number of steps. Changing the signal doesn't make it last longer. At last, the speed and tone of the sound changed using a method called time stretching. These methods help us have more training data and prevent our proposed Adaptive boosted support vector-machine random forest model from getting too specific.

3.3. Feature Extraction

Finding important information in sounds helps a lot when sorting them into different categories, especially for environmental sounds. It makes things faster and simpler by cutting down on the time it takes to compute, reducing mistakes in classification, and simplifying how the algorithm works. So, the steps are really important to pull out the most important features from the audio signals that help describe the sound around us. Different characteristics were taken out and taught using the suggested boosted SVM and Random Forest. The UrbanSound8K dataset, ESC-10, and ESC-50 were analyzed for different sound

features. Various features help understand the characteristics of sound, including Mel-spectrograms, Chromagrams, MFFCs, delta MFCCs, delta-delta MFCCs, Tonnetz representations, and Spectral Contrast.

3.4. Model Selection

Advanced machine learning methods with unique features include Random Forest and Adaptive Boosted Support Vector Machine-Random Forest. The strength of SVMs-RF, which are known for their efficiency in managing complicated, non-linear data spaces, is combined with AdaBoost, an ensemble technique that enhances performance by concentrating on difficult cases, to create Adaptive Boosted SVM-RF. Although it requires a large amount of processing power and thorough parameter adjustment, this combination is reliable for complicated classification. Yet Random Forest can be used for both regression and classification. Particularly well known for its ability to reduce overfitting and providing information about the relative importance of different features, it was especially well known for its ability to reduce overfitting that arises in individual decision trees. Although it may not work well with linear data, it was less approachable than more straightforward models.

3.5. Experimental Setup and Evaluation

A collection of N*N matrices is called a performance matrix, and it is used to assess the effectiveness of categorization models. The confusion matrix provides information on the effectiveness of the classification model and errors that it makes. In a 2*2 confusion matrix, the following 4 values are taken into account when computing:

True Positive TP: That is actual positive result model anticipated to be positive.

False Positive TN: That is actual negative result anticipated by model as negative.

True Negative FP: That is actual negative result anticipated by model as positive.

False Negative FN: That is actual positive result anticipated by model as negative.

Precision: Precision depends on how relevant the results are. It measures the proportion of overall positive results that are actually positive (true positive plus false positive) as given in Eq (1). The precision p is as follows if (TP) is a true positive and (FP) is a false positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

**Recall:** Recall is the percentage of results that are actually relevant. It is the proportion of genuinely positive results to the total of both positive and falsely negative findings as given in Eq 1.. If (TP) is actually positive and (FN) is a false negative, then remember the situation as it is.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

**F1-score:** It also goes by the name F-measure. It is the harmonic mean of precision and memory. F1 is given if P is Precision and R is Recall as given in Eq 2.

$$\text{F1} - \text{score} = \frac{2\times(\text{Pression}\times\text{Recall})}{(\text{Pression+Recall})} \quad (3)$$

**Accuracy:** The degree to which experimental data and the variables that are used to assess how effectively a classification system performed agree is referred to as accuracy as given in Eq 3. An algorithm's suitability for a particular dataset is determined by classification evaluation criteria.

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+TN+FP+FN})} \quad (4)$$

## 4. Experimental Results

Evaluates the performance of the AdaBoost SVM-RF model using the original UrbanSound8K, ESC-10, and ESC-50 datasets, as well as their enhanced versions. All experiments utilized a 10-fold cross-validation approach, ensuring that each data sample contributed to both the training and testing.

4.1. Performance of ESC-10 dataset

Applying Chromagram, Mel-spectrogram, delta MFCCs, delta-delta MFCCs, MFFCs, Spectral Contrast, and Tonnetz representation feature extraction methodologies on the original and enhanced or augmented version of ESC-10 datasets. Table 1 and Figure 4 & 5 illustrate the performance of AdaBoost SVM-RF models on both original and enhanced ESC-10. The model obtained a weighted test accuracy of 78% on original while 95% on the enhanced ESC10 coupled with the AdaBoost SVM-RF model. Furthermore, the table presents the outcomes using the F1-score, recall, and accuracy metrics. Using the AdaBoost SVM-RF model, the experimental results revealed that the Sea Waves class achieved the highest F1-score of 95%, while the baby cry class obtained the lowest F1-score of 90%. The F1-score for the

remaining classes ranged from 71% to 96%. In the Helicopter, Rooster, and Fire classes, the highest recall score was 96%, the remaining classes scores ranged from 67% to 96%. Likewise, in the Chainsaw class, the highest accuracy score was 94%.

**Table 1.** Performance (%) of Proposed model on the ESC-10

| Class | Original | | | | Augmented | | | |
|---|---|---|---|---|---|---|---|---|
| | Accur | Precisio | Recall | F1-Score | Accurac | Precision | Recall | F1-Score |
| Dog Bark | 83 | 83 | 83 | 83 | 100 | 100 | 100 | 100 |
| Rain | 50 | 44 | 50 | 47 | 87 | 94 | 87 | 90 |
| Sea Waves | 80 | 86 | 80 | 83 | 96 | 92 | 96 | 95 |
| Baby Cry | 94 | 94 | 94 | 94 | 98 | 83 | 98 | 90 |
| Clock Tick | 78 | 58 | 78 | 67 | 100 | 96 | 100 | 98 |
| Person | 70 | 64 | 70 | 67 | 100 | 95 | 100 | 97 |
| Helicopter | 89 | 80 | 89 | 84 | 96 | 100 | 96 | 98 |
| Chainsaw | 67 | 60 | 67 | 63 | 99 | 93 | 99 | 96 |
| Rooster | 85 | 92 | 85 | 88 | 88 | 100 | 88 | 93 |
| Fire Cracking | 72 | 100 | 72 | 84 | 91 | 97 | 91 | 94 |
| **Weighted** | **78** | **80** | **78** | **79** | **95** | **95** | **95** | **95** |

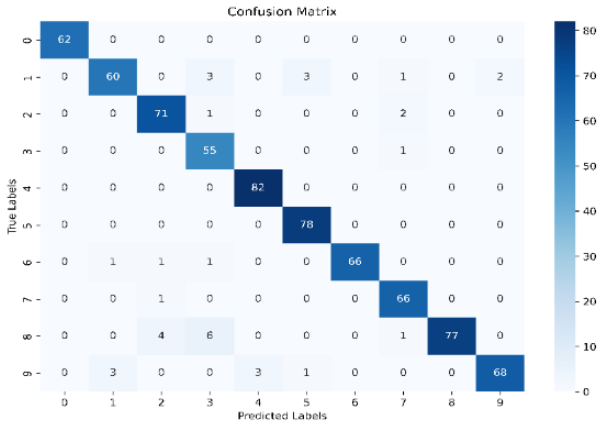



**Figure 4.** Result on original ESC10

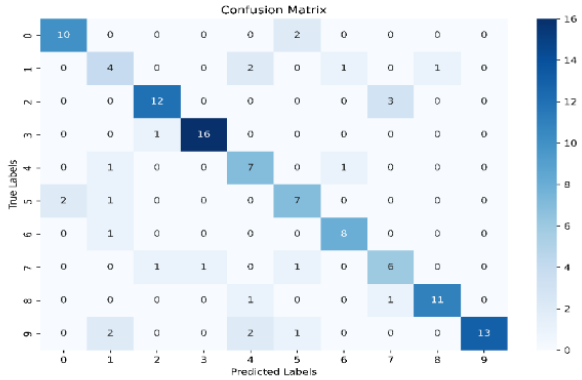



**Figure 5.** Result on augmented ESC10

The audio files underwent data enhancement techniques to reduce the risk of overfitting. After this enhancement, the AdaBoost SVM-RF model was provided with sufficient training data. The advantages of using data augmentation for training these models are demonstrated in Tables 1. The experimental results show an approximate 17% average accuracy difference between the models on the original and enhanced datasets. On the improved ESC-10 dataset, the AdaBoost SVM-RF model achieved the highest accuracy of 96%.

4.2. ESC-50 dataset

The ESC-50 dataset is significantly more complex and comprehensive. It includes a wider range of classes but offers fewer audio samples for model training compared to ESC-10, increasing the risk of

overfitting. The performance of AdaBoost SVM-RF on the original as well as enhanced ESC-50 dataset is presented in Table 2. The experimental results highlight the advantages of using AdaBoost SVM-RF which require less training time and achieved the highest accuracy of 85%. This demonstrates that ESC-50 is more susceptible to overfitting compared to ESC-10. To mitigate this, three data augmentation techniques were utilized to the ESC-50. The experimental results, shown in Table 1, indicate improvements in addressing overfitting issues. The model achieved the highest accuracy of 85%.

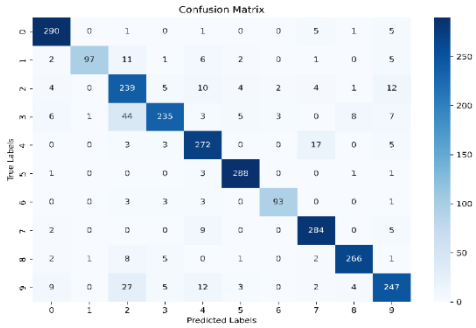
**Table 2.** Performance (%) of model on the ESC-50

| Class | Original | | | | Augmented | | | |
|---|---|---|---|---|---|---|---|---|
| | Accurac | Precisio | Recall | F1-Score | Accuracy | Precisio | Recall | F1-Score |
| Dog | 73 | 42 | 73 | 53 | 95 | 89 | 95 | 92 |
| Rooster | 77 | 56 | 77 | 65 | 93 | 88 | 93 | 90 |
| Pig | 50 | 45 | 50 | 48 | 82 | 86 | 82 | 84 |
| Cow | 92 | 61 | 92 | 73 | 95 | 95 | 95 | 95 |
| Frog | 71 | 71 | 71 | 71 | 99 | 97 | 99 | 98 |
| Cat | 42 | 42 | 42 | 42 | 80 | 80 | 80 | 80 |
| Hen | 29 | 31 | 29 | 30 | 75 | 87 | 75 | 80 |
| Insects | 42 | 33 | 42 | 37 | 79 | 97 | 79 | 87 |
| Sheep | 75 | 55 | 75 | 63 | 89 | 88 | 89 | 89 |
| Crow | 53 | 91 | 53 | 67 | 94 | 90 | 94 | 92 |
| Rain | 77 | 62 | 77 | 69 | 95 | 73 | 95 | 82 |
| Sea Waves | 81 | 76 | 81 | 79 | 94 | 86 | 94 | 90 |
| Cracking Fire | 62 | 56 | 62 | 59 | 91 | 74 | 91 | 82 |
| Crickets | 33 | 40 | 33 | 36 | 87 | 94 | 87 | 90 |
| Chirping | 67 | 60 | 67 | 63 | 93 | 87 | 93 | 90 |
| Water Drops | 20 | 60 | 20 | 30 | 85 | 93 | 85 | 88 |
| Wind | 64 | 50 | 64 | 56 | 93 | 71 | 93 | 80 |
| Pouring | 57 | 27 | 57 | 36 | 84 | 79 | 84 | 82 |
| Toilet Flush | 69 | 65 | 69 | 67 | 96 | 85 | 96 | 90 |
| Thunderstor | 100 | 40 | 100 | 57 | 87 | 86 | 87 | 86 |
| Crying Baby | 67 | 44 | 67 | 53 | 97 | 77 | 97 | 86 |
| Sneezing | 55 | 50 | 55 | 52 | 90 | 87 | 90 | 88 |
| Clapping | 88 | 70 | 88 | 78 | 87 | 84 | 87 | 86 |
| Breathing | 25 | 30 | 25 | 27 | 73 | 73 | 73 | 73 |
| Coughing | 40 | 22 | 40 | 29 | 80 | 93 | 80 | 86 |
| Footsteps | 62 | 57 | 62 | 59 | 93 | 83 | 93 | 88 |
| Laughing | 18 | 50 | 18 | 27 | 61 | 91 | 61 | 73 |
| Brushing | 60 | 40 | 60 | 48 | 89 | 71 | 89 | 79 |
| Snoring | 47 | 70 | 47 | 56 | 83 | 81 | 83 | 82 |
| Drinking | 13 | 33 | 13 | 19 | 73 | 84 | 73 | 78 |
| Door Knock | 89 | 53 | 89 | 67 | 91 | 75 | 91 | 82 |
| Mouse Click | 42 | 45 | 42 | 43 | 76 | 84 | 76 | 80 |
| Keyboard | 38 | 23 | 38 | 29 | 67 | 84 | 67 | 74 |
| Door | 23 | 100 | 23 | 38 | 62 | 88 | 62 | 73 |
| Can Opening | 33 | 33 | 33 | 33 | 97 | 83 | 97 | 90 |
| Washing | 50 | 82 | 50 | 62 | 80 | 84 | 80 | 82 |
| Vacuum | 62 | 62 | 62 | 62 | 89 | 76 | 89 | 82 |
| Clock Alarm | 64 | 90 | 64 | 75 | 93 | 92 | 93 | 93 |
| Clock Tick | 20 | 25 | 20 | 22 | 71 | 83 | 71 | 78 |
| Glass | 0 | 0 | 0 | 0 | 89 | 99 | 89 | 94 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Helicopter | 50 | 70 | 50 | 58 | 89 | 89 | 93 | 89 |
| Chainsaw | 60 | 55 | 60 | 57 | 90 | 81 | 71 | 85 |
| Siren | 40 | 50 | 40 | 44 | 90 | 90 | 89 | 90 |
| Car Horn | 19 | 75 | 19 | 30 | 67 | 96 | 89 | 79 |
| Engine | 25 | 60 | 25 | 35 | 85 | 83 | 90 | 84 |
| Train | 60 | 50 | 60 | 55 | 84 | 83 | 90 | 84 |
| Church Bells | 54 | 88 | 54 | 67 | 96 | 94 | 67 | 95 |
| Airplane | 35 | 75 | 35 | 48 | 80 | 94 | 85 | 86 |
| Fireworks | 77 | 71 | 77 | 74 | 95 | 84 | 95 | 89 |
| Handsaw | 22 | 18 | 22 | 20 | 65 | 96 | 65 | 77 |
| **Weighted** | **51** | **56** | **51** | **51** | **85** | **86** | **85** | **85** |

4.3. Urbansound8k dataset

The UrbanSound8K dataset contains relatively some class, but the repository contains as many as 8,732 audio files to train AdaBoost SVM-RF model. The result of evaluating the performance of both models on the original on the original as well as enhanced UrbanSound8K dataset is shown in Table 3. The highest accuracy 94%, was achieved by the AdaBoost SVM-RF model when using seven acoustic features: MFCCs, mel spectrogram, chromagram, delta-delta MFCCs, Tonnetz representation, and spectral contrast. The distribution of classification accuracy of each class in this dataset is presented in Figure 6 & 7. In the AdaBoost SVM-RF model on the enhanced UrbanSound8K dataset Table 4 represents the experimental results and the maximum Accuracy is 95%. It may seem that the difference in the accuracy of the classifications for the enhanced and the original version of the UrbanSound8K, ESC-10 and ESC-50 as shown in Table 4.



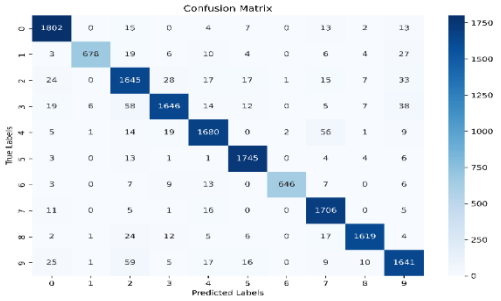




**Figure 6.** Result of model on original Us8k **Figure 7.** Result of model on augmented Us8k

**Table 3.** Performance (%) of model on the UrbanSound8K

| Class | Original | | | | Augmented | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Precisi | Recall | F1-Score | Ac | Precisio | Recall | F1-Score |
| Air Conditioner | 96 | 92 | 96 | 94 | 97 | 95 | 97 | 96 |
| Car Horn | 78 | 98 | 78 | 87 | 90 | 99 | 90 | 94 |
| Children Playing | 85 | 71 | 85 | 77 | 92 | 88 | 92 | 90 |
| Dog Bark | 75 | 91 | 75 | 83 | 91 | 95 | 91 | 93 |
| Drilling | 91 | 85 | 91 | 88 | 94 | 95 | 94 | 94 |
| Engine Idling | 98 | 95 | 98 | 96 | 98 | 97 | 98 | 97 |
| Gun Shot | 90 | 95 | 90 | 93 | 93 | 100 | 93 | 96 |
| Jackhammer | 95 | 90 | 95 | 92 | 98 | 93 | 98 | 95 |
| Siren | 93 | 95 | 93 | 94 | 96 | 98 | 96 | 97 |
| Street Music | 80 | 85 | 80 | 83 | 92 | 92 | 92 | 92 |
| **Weighted** | **88** | **89** | **88** | **88** | **94** | **95** | **94** | **94** |

**Table 4.** Performance evaluation of AdaBoost SVM-RF

| Database | Original | | Enhanced | |
|---|---|---|---|---|
| | Time | Accuracy (%) | Time | Accuracy (%) |
| ESC-10 | 75 | 78 | 102 | 95 |

| | | | | |
|---|---|---|---|---|
| ESC-50 | 354 | 51 | 455 | 85 |
| US8K | 1544 | 88 | 2210 | 94 |

4.4. Comparison of Proposed Method with Other Methods as applied on ESC

To confirm the quality of the proposed transformer-based deep learning method for ESC, the results of its use were compared with the results of similar models, using the same datasets as shown in Table 4.5. We have chosen six articles published in the last few years that investigated ESC and compared it to the ones cited here as [22, 25,27]. In study [22], they presented a novel hybrid feature generation technique for combining both the textual and statistical characteristics from the sound files for the ESC. Extracted feature set comprised 1D Local Binary Pattern and feature 1D Ternary Pattern like mean and median etc. were fed in to a third-order SVM that provided a accuracy of 91%. In Study [25], the author used data augmentation and an enhanced CNN to extract the Mel-Spectrogram of the audio samples. To overcome the signal variations, this study also adopted transfer learning and data augmentation mechanisms, and the classification accuracy of the model achieved 93%. The initial training stage of the proposed framework is achieved by using the automated parameters tuning with 3% on the Us8K datasets. In another study [27], authors have discussed about deep convolutional neural network and denoising of ESC using CNN. Some of the models they used include RNN and LSTM for the direct extraction of features, followed by a support vector machine for classification.

## 5. Conclusions

This study presents an adaptive boosted support vector machine-random forest (AdaBoost SVM-RF) collaborative model for ESC (Environmental Sound Classification). The model includes a new combination of features, MFCCs, mel spectrogram, chromagram, delta-delta MFCCs, Tonnetz representation, and spectral contrast, normalized pitch frequency to obtain spectral characteristics, statistical properties, and frequency information from the audio data. Through this fusion of multimodal information, the model discriminative ability is highly improved. The adaptation of different machine learning models included in the framework AdaBoost SVM-RF deals with the robustness and generalization ability to minimize the training time and increase the classification rate under limited computational power. In the same regard, time stretch, pitch tuning, and white noise are employed to reduce overfitting especially for small datasets with a limited number of audio samples per class. The model performance was evaluated on three datasets: ESC-10, ESC-50, and UrbanSound8K databases. The AdaBoost SVM-RF classification rates of 95\% on ESC-10 dataset, 85\% on ESC-50 dataset and 94\% on the urban sound dataset. Specifically, the performance was enhanced in terms of accuracy, precision, recall and F1-score over the various classes and datasets, which supports the contribution of the fused features. AdaBoost SVM-RF as the classifier integrated with feature extraction methods and data augmentation ensures a reliable classification of environmental sounds. Potential improvement in ESC is illustrated in the study by its potential uses in areas like monitoring noise in cities, observing wildlife, and in smart city solutions. The outcomes of this research allow for the further development of the ESC methodologies as well as improvements in the methods for environmental sound analysis and classification.

**References**


1. Hafiz Yasir Ghafoor, Rashid Jahangir, Arfan Jaffar, Roobaea Alroobaea, Oumaima Saidani, and Fatimah Alhayan. Sensors-based human activity recognition using hybrid features and deep capsule network. *IEEE Sensors Journal*, 24(14):23129–23139, 2024.
2. Rashid Jahangir, Mohammed Alreshoodi, and Fawaz Khaled Alarfaj. Spectrogram features-based automatic speaker identification for smart services. *Applied Artificial Intelligence*, 39(1):2459476, 2025.
3. Rashid Jahangir, Ying Wah Teh, Uzair Ishtiaq, Ghulam Mujtaba, and Henry Friday Nweke. Automatic speaker identification through robust time domain features and hierarchical classification approach. *In Proceedings of the international conference on data processing and applications*, pages 34– 38, 2018.
4. Jahangir Rashid, Ying Wah Teh, Hanif Faiqa, and Mujtaba Ghulam. Correction to: Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*, 80(16):23813–23813, 2021.
5. Rashid Jahangir, Muhammad Asif Nauman, Roobaea Alroobaea, Jasem Almotiri, Muhammad Mohsin Malik, and Sabah M Alzahrani. Deep learning-based environmental sound classification using feature fusion and data enhancement. *Computers, Materials & Continua*, 75(1), 2023.
6. S Chandrakala and SL Jayalakshmi. Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Computing Surveys (CSUR)*, 52(3):1–34, 2019.
7. Wei Wang, Fatjon Seraj, Nirvana Meratnia, and Paul JM Havinga. Privacy-aware environmental sound classification for indoor human activity recognition. *In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 36–44, 2019.
8. Ozkan Inik. CNN hyper-parameter optimization for environmental sound classification. *Applied Acoustics*, 202:109168, 2023.
9. Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang. Environmental sound classification with parallel temporal-spectral attention. *arXiv preprint* arXiv:1912.06808, 2019.
10. Muhammad Zohaib Anwar, Zeeshan Kaleem, and Abbas Jamalipour. Ma chine learning inspired sound-based amateur drone detection for public safety applications. *IEEE Transactions on Vehicular Technology*, 68(3):2526–2534, 2019.
11. Biyun Ding. Low-complexity acoustic scene classification using simple cnn. Technical report, Tech. Rep., DCASE 2019 Challenge, 2019.
12. Mohammad Esmaeilpour, Patrick Cardinal, and Alessandro Lameiras Ko erich. From sound representation to model robustness. *arXiv preprint* arXiv:2007.13703, 2020.
13. Zohaib Mushtaq and Shun-Feng Su. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167:107389, 2020.
14. Yu-Kai Lin, Mu-Chun Su, and Yi-Zeng Hsieh. The application and improvement of deep neural networks in environmental sound recognition. *Applied Sciences*, 10(17):5965, 2020.
15. Heng Li, Qing Zhang, Xianrong Qin, and Sun Yuantao. Raw vibration signal pattern recognition with automatic hyper-parameter-optimized convolutional neural network for bearing fault diagnosis. Proceedings of the Institution of Mechanical Engineers, Part C: *Journal of Mechanical Engineering Science*, 234(1):343–360, 2020.
16. Mohammad Esmaeilpour, Patrick Cardinal, and Alessandro Lameiras Koerich. Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network. *Applied Soft Computing*, 86:105912, 2020.
17. Lancelot Lhoest, Mimoun Lamrini, Jurgen Vandendriessche, Nick Wouters, Bruno da Silva, Mohamed Yassin Chkouri, and Abdellah Touhafi. Mo saic: A classical machine learning multi-classifier based approach against deep learning classifiers for embedded sound classification. *Applied Sciences*, 11(18):8394, 2021.
18. David Elliott, Carlos E Otero, Steven Wyatt, and Evan Martino. Tiny trans formers for environmental sound classification at the edge. *arXiv preprint* arXiv:2103.12157, 2021.
19. Dipro Pramanick, Haaris Ansar, Hemant Kumar, S Pranav, Richa Tengshe, and Binish Fatimah. Deep learning based urban sound classification and am bulance siren detector using spectrogram. *In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2021.
20. Ram CMC Shekar, Chelzy Belitz, and John HL Hansen. Development of cnn based cochlear implant and normal hearing sound recognition models using natural and auralized environmental audio. *In 2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 728–733. IEEE, 2021.

21. Jurgen Vandendriessche, Nick Wouters, Bruno da Silva, Mimoun Lamrini, Mohamed Yassin Chkouri, and Abdellah Touhafi. Environmental sound recognition on embedded systems. *Electronics*, 10(21):2622, 2021.
22. Yusuke Misumi, Shigeru Miyagawa, Daisuke Yoshioka, Satoshi Kainuma, Takuji Kawamura, Ai Kawamura, Yuichi Maruyama, Takayoshi Ueno, Koichi Toda, Hidetsugu Asanoi, et al. Prediction of aortic valve regurgitation after continuous flow left ventricular assist device implantation using artificial intelligence trained on acoustic spectra. *Journal of Artificial Organs*, 24:164–172, 2021.
23. Achyut Mani Tripathi and Aakansha Mishra. Adv-esc: Adversarial attack datasets for an environmental sound classification. Applied Acoustics, 185:108437, 2022.
24. Sharayu Kawale, Dharmesh Dhabliya, and Ganesh Yenurkar. Analysis and simulation of sound classification system using machine learning techniques. *In 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS)*, pages 407–412. IEEE, 2022.
25. Jinming Guo, Chuankun Li, Zepeng Sun, Jian Li, and Pan Wang. A deep attention model for environmental sound classification from multi-feature data. *Applied Sciences*, 12(12):5988, 2022.
26. Grach Mkrtchian and Yury Furletov. Classification of environmental sounds using neural networks. *In 2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)*, pages 1–4. IEEE, 2022.
27. Asadullah Ashurov, Yi Zhou, Liming Shi, Yu Zhao, and Hongqing Liu. Environmental sound classification based on transfer-learning techniques with multiple optimizers. *Electronics*, 11(15):2279, 2022.
28. Asadullah Ashurov, Zhou Yi, Hongqing Liu, Zhao Yu, and Manhai Li. Concatenation-based pre-trained convolutional neural networks using atten tion mechanism for environmental sound classification. *Applied Acoustics*, 216:109759, 2024.
29. P Priyanka Jesudhas and P Vanaja Ranjan. A novel approach to build a low complexity smart sound recognition system for domestic environment. *Applied Acoustics*, 221:110028, 2024.