University of Baghdad
College of Engineering
JOURNAL OF ENGINEERING

Journal of Engineering
journal homepage: www.jcoeng.edu.iq
Volume 32 Number 8 August 2026

Environmental Sound Classification Using a Hybrid SVM-CNN Approach

Nisreen Talib Abdulhusein, Basheera M. Mahmmod*

Department of Computer Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq

ABSTRACT

Classifying environmental sounds is considered a challenging task because of the complex nature of acoustic signals. This classification involves identifying the acoustic flow associated with different sounds in the environment. Several factors affect it, such as the distances between the source and the destination, voice interference, and the huge number of sound sources. These problems have been addressed in this work, where a hybrid deep learning technique is proposed to improve the classification of sounds, focusing on fundamental audio features which are accurate in these tasks. The proposed hybrid classification framework combines a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). The SVM model uses feature extraction based on the orthogonal transformation (DCT), and the CNN model uses machine learning to represent data from Log-Mel spectral diagrams. Noisy sound signals are generated using the TIMIT and NOISEX-92 datasets by combining the clean signal with various types of environmental sounds, resulting in 17 categories, including the clean signal category. A soft voting strategy is then used to enhance overall decision-making. It calculates the weighted average probability for each class across all models and selects the class that has the highest average probability as the final output with more robust and reliable classification. The experimental results prove that the proposed model achieves high performance compared to current ways and individual models, reaching an accuracy of 99.28% with improvements in F1 score, accuracy and recall. The results obtained prove the efficiency of integrating ML and DL methods for classifying environmental sounds.

Keywords: CNN, Deep learning, Environmental sound classification, Machine learning, SVM, Soft voting.

1. INTRODUCTION

Handling different types of environmental sounds has become a major concern in modern urban environments due to their impact on quality of life, applications, and the performance of communication systems **(Meedeniya et al., 2023)**. In particular, environmental noise can significantly affect the quality and intelligibility of transmitted audio signals, especially

*Corresponding author
Peer review under the responsibility of University of Baghdad.
https://doi.org/10.31026/j.eng.2026.08.11

Article received: 22/04/2026
Article revised: 18/06/2026
Article accepted: 26/06/2026
Article published: 01/08/2026

speech signals **(Nasir and Abdulmohsin, 2024; Hussein et al., 2023)**. This makes Environmental Sound Classification (ESC) a crucial task in signal processing applications, especially in speech processing such as speech enhancement **(Yousif and Mahmmod, 2026)** and Automatic Speech Recognition (ASR) **(Ibraheem and Shihab, 2024; Chu et al., 2023).** Substantial progress has been made in improving the accuracy of sound classification systems based on the quick advancement of machine learning and deep learning technologies **(Beritelli et al., 2023)**.

However, classifying environmental sounds, such as music, video clips, speakers, non-speech and environmental noise sounds, is considered challenging due to the wide diversity of sound sources and their various acoustic characteristics **(Bansal and Garg, 2022)**. The sources of these sounds in general can be classified into two main types: natural sources such as rain, wind, and birdsong, as well as artificial sources like traffic noise, alarms, and machinery **(Nogueira et al., 2022)**. These audio signals are often affected by background scenarios, and they are often irregular, involving overlap between different sounds and variations in sound intensity, which increases the complexity of the classification process and requires advanced techniques for feature extraction using robust learning models. Despite these challenges, research into the classification of environmental acoustic sound continues and is of great importance in many practical fields, such as the medical field for patient monitoring **(Zhang et al., 2024; Aji et al., 2024)**, smart cities for monitoring traffic and accidents **(Ilahi et al., 2025; Walden et al., 2022)**, security systems for detecting abnormal events **(Chan and Eric, 2010; Iqbal et al., 2025)**, and wildlife monitoring and environmental studies **(Hasan, 2022; Islam and Ali, 2024)**. Many methods have been proposed in the literature for classifying environmental sound, and different machine learning methods have been proposed for this task, such as Support Vector Machines (SVMs) **(Blanco et al., 2022)**, K-Nearest Neighbors (KNN), and Random Forest (RF) **(Alsouda et al., 2019)**. These methods have been widely approved due to their ability to perform well when used with manually extracted acoustic features. Features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectral representations are used to distinguish between different types of sound **(Ali et al., 2022).** In recent years, deep learning techniques **(Zaman et al., 2023)**, particularly convolutional neural networks (CNNs) **(Demir et al., 2020)**, have acquired more attention for their ability to automatically learn features from frequency-time representations such as the acoustic spectrum (spectrogram). Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have also been used to process acoustic signals containing sequential time information **(Sankavi et al., 2024).**

More recently, hybrid methods have been introduced that combine multiple classifiers to exploit the advantages of each model **(Bansal and Garg, 2023)**. These methods aim to improve classification accuracy under complex acoustical conditions **(Akter et al., 2025)**. In limited data scenarios, these systems utilize the best of both worlds: the ability of deep learning to learn hierarchical features and the efficiency of traditional machine learning in limited data scenarios. Also, by combining predictions from multiple models (clustering), these hybrid methods can improve generalization and reduce variance **(Akter et al., 2025)**. Mostly, these models are characterized by high computational efficiency, making them suitable for edge device applications and real-time audio scene classification **(Goulão et al., 2024)**. However, improving accuracy in noisy or reverberant environments is still a critical issue. This research suggests a hybrid system for classifying environmental sound based on the integration of SVM and CNN classifiers called the Hybrid Classification Model (HCM). The SVM classifier uses the extracted features from the time signal using Orthogonal transform

features, while the CNN classifier relies on the Log-Mel Spectrogram representation to extract features from the time-frequency domain. The outputs of the two classifiers are combined using soft voting to get the final decision, i.e., using soft voting to improve overall decision-making. The results prove that this hybrid model outperforms some related models and each classifier individually, achieving an accuracy of approximately 99%.

Unlike conventional environmental sound classification systems that depend on a single feature representation and classifier, the proposed approach combines complementary information extracted from orthogonal transform-based features for SVM classifier and log-Mel spectrogram features for CNN classifier. Furthermore, a soft-voting fusion strategy is employed to integrate the decisions of the SVM and CNN classifiers. The proposed framework performs synchronized classification using identical audio segments and achieves robust performance across several environmental sound classes and different SNR conditions.

These characteristics distinguish the proposed method from existing environmental sound classification approaches.

The main contributions of this work are listed below:

1. To develop a new model for sound classification using two powerful Neural Networks with stacked features and explore the effectiveness of different noise features like Orthogonal Polynomial Transform features.
2. To evaluate and compare the performance of the proposed approach with the traditional classifier method based on many sound types.
3. To explore and analyze the impact of these techniques on enhancing the classification performance and efficiency of the proposed classification methodology.

2. RELATED WORKS

Environmental Sound Classification (ESC) has witnessed widespread interest in recent years, with numerous search methods proposed to raise classification accuracy. In general, these methods can be classified into three main groups: traditional machine learning-based methods, deep learning-based methods, and hybrid architectures that combine more than one model.

2.1 Machine Learning Methods

In the literature, classification of environmental background sounds is used in many sound applications and platforms, such as voice recognition and speech enhancement. Different studies have used traditional ML methods for ECS such as SVM, KNN, Random Forest, and Bagging.

SVMs are one of the methods that have been widely used due to their strong classification capability. For example, **(Abdoune et al., 2024)** applied SVM with MFCC features and achieved high accuracy, reaching 100% for a small number of classes. However, the performance decreased when the number of classes increased, reaching 95.19%, which signifies limitations in complex conditions.

In another study **(Malhotra, 2023)**, several classifiers, including SVM, KNN, and Random Forest, used MFCC features for urban noise classification. The results showed that all models reached high accuracy, ranging between 95% and 100%. Though the high classification accuracy is stated in the study, many limitations can be found. First, the dataset used is relatively small, containing a limited number of audio recordings, which may affect the generalization capability of the model. Second, the number of sound classes is four with only,

making the classification task less complex compared to real-world conditions, which contain a lot of types of noise.

Many other studies have investigated the use of machine learning techniques for ESC. For instance, **(Albaji et al., 2023)** suggested using various classifiers, such as SVM, KNN, decision tree, logistic regression, and random forest (RF). The study utilized a dataset consisting of 873 samples of environmental sounds classified into 16 different noise types. The results showed that the random forest classifier achieved the best performance among the other tested models.

2.2 Deep Learning Methods

Modern deep learning approaches such as DNN, CNN, LSTM, and BiLSTM have been recently applied to address the non-linear and unstructured nature of sound in the real world, which are used in ESC tasks. For example, **(Sankavi et al., 2024)** proposed a deep learning-based approach using either convolutional neural networks (CNN) 1D and 2D, or long short-term memory (LSTM) models based on two types of features: spectrogram or autocorrelation features. The study assessed the models on noisy speech signals corrupted by different types of noise and reported high classification performance. The spectrogram-based LSTM model realizes up to 99.6% accuracy. In addition, the proposed technique focused on reducing computational complexity and making it suitable for real-time applications.

Another study, **(Beritelli et al., 2023)** suggested CNN as a method of classification. This technique combined many types of datasets to make a diverse training dataset, using 14 types of noise and treated raw audio signals directly rather than depending on traditional handcrafted features. The results proved high classification accuracy, reaching 96.93% and increasing to 100% with post-processing techniques. In addition, **(Souadek and Guellil, 2025)** proposed a CNN-based model combined with Short-Time Fourier Transform (STFT) features for urban sound classification. The model was trained on multiple datasets and reached an accuracy of around 95%.

2.3 Hybrid Methods

Recently, there has been a movement towards using hybrid approaches that combine more than one model to benefit from the advantages of each. In order to accurately classify environmental sounds, which is considered a challenging task due to the limited data, diversity of sounds, and the difficulty of extracting features, several studies based on these concepts were conducted as mentioned below. **(Akter et al., 2025)** proposed a hybrid model combining CNN and LSTM to improve classification accuracy, using MFCC and Mel spectrogram features. The results displayed that this model beats traditional methods, indicating that hybrid models are more effective in this field. Another study, **(Hasnain et al., 2025),** suggested an adaptive boosted support vector machine-random forest (AdaBoost SVM-RF) for ESC as a cooperative model. The proposed model integrates two classifiers, namely SVM and Random Forest. A variety of features, including MFCC, Mel-spectrogram, and spectral properties, are used to enhance classification performance. The results obtained proved that this work outperformed different models, with accuracies of 94%, 85%, and 95% across various databases in the ESC field.

In addition, another study concentrated on bioacoustics signal classification using deep learning approaches represented by **(Hu et al., 2024)**. This work proposed a hybrid model that integrates CNN, bi-directional long short-term memory (BiLSTM), and an attention mechanism to enhance feature extraction and classification performance. The model was

assessed on biologically sound datasets and achieved high performance, reaching an accuracy of 98.35%, indicating strong generalization ability.

Table 1 provides a comprehensive summary with a comparative analysis of various existing works to get a better understanding about the differences among current methods. The table illustrates different types of methods, including ML, DL, and hybrid models, along with specific characteristics, strengths, datasets, and main limitations.

Table 1. A comparative analysis of some existing studies

Ref.	Details	Accuracy	Strength	Limitation
(Abdoune et al., 2024)	**Type**: ML **Model:** SVM **Features:** MFCC **Dataset:** Freesound **No. of noise:** 7	95.19%	Simple and low computational complexity.	Performance decreases as the number of classes increases.
(Malhotra, 2023)	**Type:** ML **Model:** SVM, KNN, RF **Features:** MFCC (primary), ZCR, MPEG-7. **Dataset:** Custom dataset (44 recordings) **No. of noise:** 4	95%	high accuracy, simple implementation.	Small dataset, limited classes.
(Sankavi et al., 2024)	**Type:** DL **Model:** CNN(1D)/CNN(2D)/ LSTM **Features:** Autocorrelation Function (ACF) or Spectrogram. **Dataset:** Malayalam + LJ Speech +MS-SNSD. **No. of noise:**6	99.60% (LSTM+ Spectrogram)	High accuracy, low computational cost.	Limited noise types, higher processing time in LSTM.
(Beritelli et al., 2023)	**Type:** DL **Model:** CNN(1D) **Features:** Raw audio (no feature extraction). **Dataset:** UrbanSound +Demand +Noisex-92. **No. of noise:**14	96.93%, 100% (with post-processing)	No pre-processing, high accuracy.	Depends on large data and needs post-processing.
(Hu et al., 2024)	**Type:** hybrid **Model:** CNN + BiLSTM + Attention **Features:** MFCC, Mel-spectrogram. **Dataset:** Custom Dataset **No. of noise:** 10	98.35%	Very high accuracy, Combines spatial, temporal, and attentional information, Powerful in complex noise environments.	Requires large datasets (despite augmentation), High computational cost, Designed for bioacoustics applications (not general ESC).

Table 1 shows that ESC techniques change among ML, DL, and hybrid models. Where, CNN model achieves higher accuracy through automatic feature learning, traditional machine

learning methods like SVM depend on manually extracted features. Through combining numerous strategies, hybrid models proved to have reliable performance. However, challenges continue, like the need for large datasets and extreme computational costs, and the advance of more efficient models.

3. METHODOLOGY

In this section, we will describe the proposed methodology for classifying environmental sounds. The proposed architecture is termed a Hybrid Classification Model (HCM), which has achieved state-of-the-art results. It is established based on a hybrid of two main stages: a traditional machine learning pipeline (SVM) classifier combined in series with a modern deep learning approach (CNN) classifier. This combination used soft voting to improve overall decision-making. Each stage is used to leverage different features of sound environmental signals, consequently enhancing classification efficiency. The classification relies on categorizing environmental sounds into different classes. These classes include speech signals that represent spoken language (clean speech signal) and a babble signal. The other category represents the non-speech sounds, which in turn have two types: machine sounds and environmental noise. Machine sounds include the sounds produced by vehicles and mechanical systems, such as cars, F-16 fighter jets, Factory 1 and 2, and Buccaneer 1 and 2 aircraft. While environmental noise includes background and ambient sounds, such as white, pink noise, Hall, and Airport noise.
This work classifies these different environmental sounds by extracting their features from diverse sound samples, then pre-processing them to train the proposed hybrid model to classify them accurately. The stages of the proposed ESC (HCM) are described in the following sections.

3.1 Dataset Description

The proposed environmental sound classification system was assessed using a dataset consisting of clean speech and environmental sound recordings. The environmental sound recordings used in this study are obtained from the NOISEX-92 dataset **(Hussein et al., 2023)**, while the clean speech signals are got from the TIMIT speech corpus. The noisy speech signals are generated by adding environmental noise to clean speech signals at predefined signal-to-noise ratio (SNR) levels (-10, -5, 0, 5, and 10 dB). The generated noisy signal is got according to Eq. (1), which represents:

$$x(n) = s(n) + d(n) \tag{1}$$

where $(s(n))$ denotes the clean speech signal, $(d(n))$ is the noise signal, and $(x(n))$ is the resulting noisy speech signal, which is then segmented and used for training and evaluation. The signal-to-noise ratio (SNR) is used to control the noise level added to the clean speech signals. It is defined as the ratio between the power of the clean speech signal and the power of the noise signal, and can be expressed in Eq. (2) as follows:

$$\text{SNR (dB)} = 10 \log_{10}\left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} d^2(n)}\right) \tag{2}$$

The dataset consists of 2,768 audio files distributed over 17 classes. including 16 environmental noise classes and one clean speech class. Most sound classes contain 32 audio files at each SNR level (-10, -5, 0, 5, and 10 dB), resulting in 160 speech files per class. Where,

the Leopard2 class contains 48 audio files per SNR level (240 files in total), while the clean speech class contains 128 files as shown in **Table 2**.

Table 2. The representation of the sound classes used in this work.

Class name	SNR Levels (dB)	Segments per SNR level	Total Number of Samples
Airport	-10, -5,0,5,10	32	160
Babble	-10, -5,0,5,10	32	160
Buccaner1	-10, -5,0,5,10	32	160
Buccaner2	-10, -5,0,5,10	32	160
Car	-10, -5,0,5,10	32	160
Destroyerengine	-10, -5,0,5,10	32	160
Destroyerops	-10, -5,0,5,10	32	160
F16	-10, -5,0,5,10	32	160
Factory1	-10, -5,0,5,10	32	160
Factory2	-10, -5,0,5,10	32	160
Hall	-10, -5,0,5,10	32	160
M109	-10, -5,0,5,10	32	160
Pink	-10, -5,0,5,10	32	160
SpeechS	-10, -5,0,5,10	32	160
White	-10, -5,0,5,10	32	160
Leopard2	-10, -5,0,5,10	48	240
Clean	/	/	128

The original audio recordings have an average duration of approximately 3 seconds. For processing and classification purposes, each audio file was segmented into fixed-length segments consisting of 2560 samples (0.16 s at a sampling frequency of 16 kHz). Furthermore, the dataset was divided using a stratified sampling strategy implemented through MATLAB's "splitEachLabel" function. Specifically, 80% of the files (2,214 samples) were used for training, 10% (277 samples) for validation, and 10% (277 samples) for testing. A fixed random seed (rng(42)) is employed to ensure reproducibility.

3.2 SVM-Based Classifier Design

SVM-based classifier is used as a machine learning method for ESC. In this work, the features are extracted manually (Handcrafted features derived from the audio signal using predefined mathematical formulas, rather than being learned automatically by the model) from the audio signals and used as input to the classifier. The orthogonal transform features are extracted from the time-domain signal using the Discrete Cosine Transform (DCT). First, the audio files were downloaded from the root directory, where each subfolder represented one sound type. The type of label of each audio sample was automatically specified according to the first folder name after the root directory.

All audio signals were resampled to a target sampling frequency f_s to ensure standard representation. Each signal was divided into N frames, and each frame had a length of L samples. Accordingly, the total segment length is defined in Eq. (3):

$$N_s = N \cdot L \tag{3}$$

This fixed-length representation was used to format the input structure for both SVM and CNN branches. The dataset is divided into training, validation, and testing subsets to ensure a proper balanced distribution of samples across all classes. The specific split ratios are presented in the experimental results section. Orthogonal transform features were computed using DCT for the extraction process. A DCT matrix was created of the same size as the frame length ($L * L$), and the first K coefficients are selected to form the orthogonal transform matrix, where K represents the moment order. Each audio segment was split into N frames, and the orthogonal transform features are calculated for each frame. The k-th moment (Yassen and Abdulhussain, 2025) is defined in Eq. (4) as:

$$m_k = \sum_{n=0}^{L-1} x(n)\varphi_k(n) \tag{4}$$

where $x(n)$ represents the input audio frame, $\varphi_k(n)$ denotes the DCT basis function, k is the moment index ranging from 0 to K-1, L is the frame length. And n is the sample index within each frame, ranging from 0 to $L - 1$. The DCT basis function (Khayam, 2003) is defined in Eq. (5) as:

$$\varphi_k(n) = \sqrt{2/L}\cos\left(\frac{\pi(2n+1)k}{2L}\right) \tag{5}$$

For each audio sample, the extracted frame-level features are then concatenated to construct the final feature vector. After feature extraction, to standardize the feature values, a Z-score normalization is applied. This normalization is defined in Eq. (6) as:

$$z = \frac{x-\mu}{\sigma} \tag{6}$$

where x represents the original feature value, μ is the mean, and σ is the standard deviation. The normalization parameters are estimated from the training set only and then applied to the validation set and testing set. In order to ensure equal scaling of all features before training the SVM model, the Z-score normalization step was necessary.

A multi-class SVM model was trained using the Error-Correcting Output Codes (ECOC) framework, which is used for classification. The model is trained to discriminate between different sound classes depending on the extracted feature representations. The specific composition of the SVM model is determined experimentally, and it is presented in the experimental results section. The trained SVM model is then used to predict the class labels of the validation and test sets. The processing pipeline of the SVM-based classifier is shown in **Fig. 1**.

Regarding the SVM classifier, the hyperparameters used in this study have been explicitly reported. Additional experiments were conducted to examine the effect of different BoxConstraint (C) values. The results showed that C = 1, 10, and 100 achieved the same highest classification accuracy (99.22%), while C = 0.1 resulted in slightly lower performance (98.44%). The ECOC-SVM classifier is applied using an RBF kernel with a BoxConstraint value of 1 and KernelScale set to “auto”, which provided satisfactory performance during many experiments.

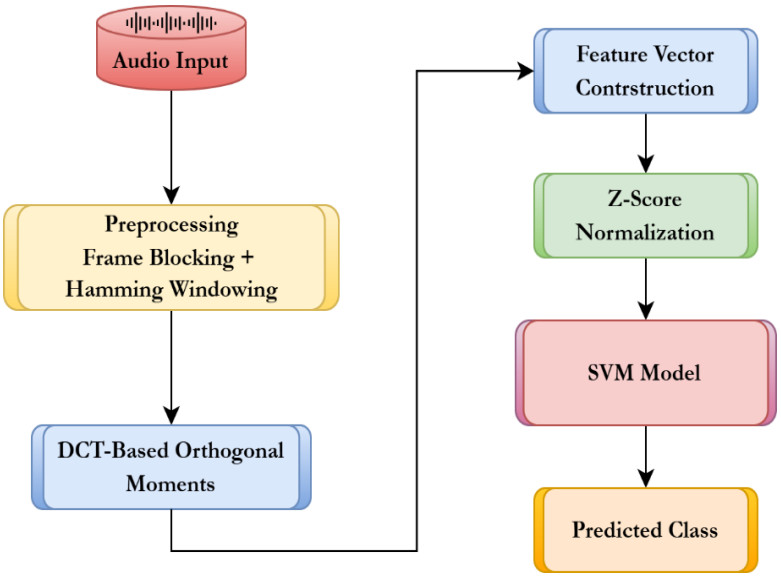

Figure 1. Flowchart for designing a classifier based on the SVM algorithm.

3.3 CNN-Based Classifier

CNN-based classifier is used as a deep learning method for environmental sound classification **(Piczak, 2015)**. This method is designed to learn discriminative features automatically from time–frequency representations of the input audio signals. So that, before being used as input to CNN, each audio segment was transformed into a fixed-size Log-Mel spectrogram.

The resulting representation has a fixed size of $(M * N)$, where M denotes the number of Mel frequency bands is set to 64, ensuring a consistent input size for the CNN model. In order to make the sound signal suitable for deep learning features, this transformation keeps both spectral and temporal information of this signal. After the step of spectrogram generation, the transformed data were joined with their corresponding class labels to construct the final datasets used for training and validation in the CNN model. In this way, to enable supervised learning of environmental sound classes, the CNN branch received paired inputs consisting of Log-Mel spectrograms and their associated labels.

The input to the CNN is a two-dimensional representation of the size of $M * N * 1$, corresponding to the number of Mel bands, the number of frames, and the channel depth. where the channel depth is set to 1, since the Log-Mel spectrogram is treated as a single-channel image. The CNN architecture consists of multiple convolutional layers, specifically three convolutional blocks, each including a convolutional layer followed by batch normalization and a ReLU activation function for non-linearity. The convolutional layers are responsible for detecting patterns in both the time and frequency dimensions of the spectrogram. Max-pooling layers were applied to reduce the spatial dimensions and improve computational performance. Then, to reduce the feature maps into a compact representation and summarize features, a global average pooling layer is applied. Before the final classification stage, a dropout layer is also incorporated to reduce overfitting. Finally, a fully connected layer maps the extracted features into the number of output classes for the final decision, followed by a softmax layer to produce class probabilities. The network is trained using an adaptive optimization algorithm with experimental parameters. The specific training settings are detailed in the experimental results section. The training data are shuffled during the learning stage to enhance generalization. To generate prediction scores

for the test dataset, the final trained network was then used.
The complete CNN network architecture and training configuration are presented in **Table 3** to improve the reproducibility of the proposed CNN classifier. The network is composed of three convolutional blocks followed by global average pooling, dropout, a fully connected layer, and a softmax classification layer.

Table 3. Network architecture and training configuration of CNN.

Layer	Configuration
Input Layer	(64×10×1) log-Mel spectrogram
Conv2D-1	32 filters, kernel size 3×3
Max Pooling-1	2×1
Conv2D-2	64 filters, kernel size 3×3
Max Pooling-2	2×1
Conv2D-3	128 filters, kernel size 3×3
Global Average Pooling	Converts feature maps into a feature vector
Dropout	0.2
Fully Connected Layer	Classes Num. (17)
Softmax Layer	produce class probabilities
Classification Layer	Final prediction of class
Learning rate	0.001

The CNN is trained using Adam optimizer for 25 epochs with a mini-batch size of 128. The validation set was used during training to observe model performance and reduce overfitting. The dropout value was set to 0.2 to improve generalization. The processing pipeline of the CNN-based classifier is shown in **Fig. 2**.

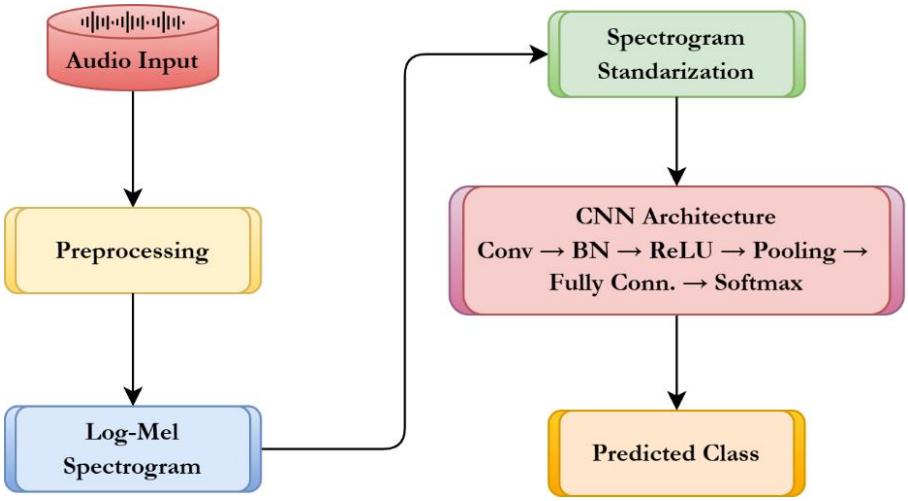

Figure 2. Flowchart for designing a classifier based on the CNN algorithm

3.4 The Hybrid-Based Classifier (HCM)

The hybrid classifier was advanced by integrating the outputs of both the SVM-based classifier and the CNN-based classifier using a soft voting ensemble approach. The purpose of this combination was to utilize the strengths of the classification of both models, where the SVM branch depends on manually orthogonal transform features, and the CNN branch

utilizes Log-Mel spectrograms to learn discriminative representations automatically from it. During the testing stage, the SVM model produced a probability score using the normalized feature vectors of the test set. In the same manner, the CNN model generated class probability scores for the same test samples based on their Log-Mel spectrogram representations. These probability outputs are then merged using a weighted averaging scheme. The ensemble decision is computed in Eq. (7) as:

$$S_{ens} = w_{SVM} \cdot S_{SVM} + w_{CNN} \cdot S_{CNN} \quad (7)$$

Where S_{SVM}, S_{CNN} denote the class probability scores obtained from the SVM and CNN models, respectively, and w_{SVM}, w_{CNN} represent their corresponding weights. Soft voting averages the probability scores for each class and selects the class with the highest average probability. The specific values of these weights are provided in the results section. The final predicted class is obtained by selecting the class with the maximum ensemble score. The classification accuracies are calculated on the test set of the individual SVM model, the CNN model, and the ensemble model is used to evaluate the efficiency of the HCM classifier. This comparison was used to determine whether integrating both classifiers could improve the overall classification performance. The soft voting mechanism was enhanced because it depends on the confidence level of each classifier rather than relying only on the final class label. This makes the ensemble decision more powerful, especially when one classifier assigns a high probability to the correct class while the other classifier has less certainty. By integrating both decision sources, the hybrid model can achieve more reliable and improved classification performance. The overall framework of the hybrid proposed classifier is shown in **Fig. 3**.

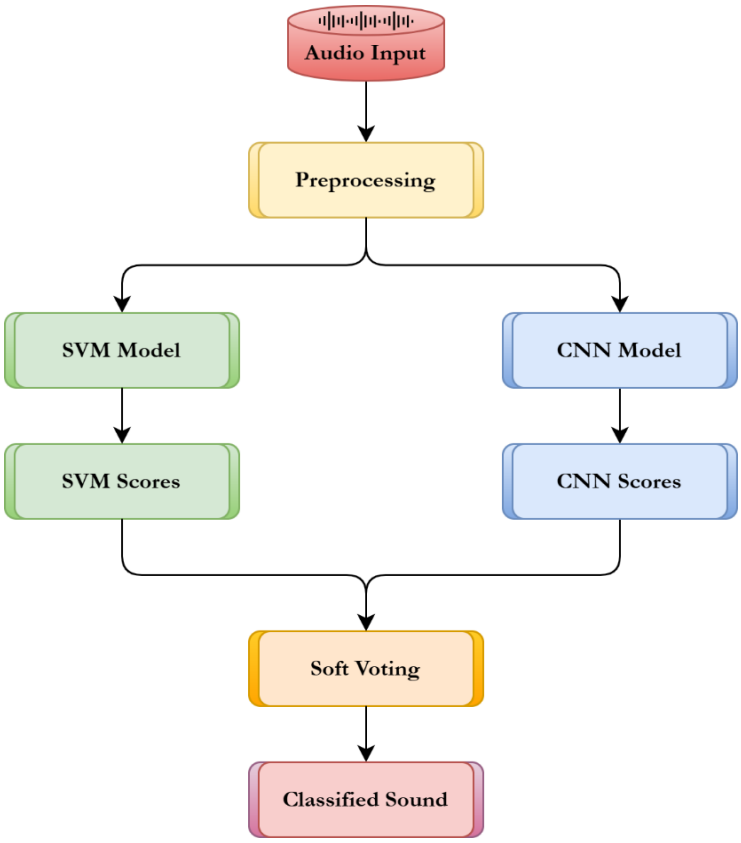

Figure 3. Flow diagram of the proposed (HCM) classifier.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the results of the experiments conducted during the feature selection and classifier performance evaluation phases. Selecting the best features, according to the methods described in the methodology section, resulted in the highest cross-validation accuracy and the best selected sound class. The experimental results of the proposed hybrid model are presented and compared with those of the traditional SVM and CNN classifiers to demonstrate its improved performance.

4.1 Experimental Setup

The audio signals are used at a sampling frequency of 16 kHz (f_s). Each signal is segmented into frames (N), where N is changed to one of three values: 10, 20, or 30, with 10 being found to be the best, with a frame length of samples (L) experiment, the values 129, 256, 384, 512, which found 256 is the best, resulting in a total segment length of 2560 samples (N_s).
For the SVM-based classifier, the orthogonal transform order (K) is set to 60 after experimenting with the values 50, 60 and 70. The value of 60 represents a balance between taking important information from the signal and not introducing unnecessary details, as lower values such as 50 may result in the loss of some information, while higher values such as 70 may introduce noise or increase complexity. The SVM model is trained using an RBF kernel because of possible nonlinear separation among the environmental noise classes. The kernel scale was set to "auto", allowing MATLAB to estimate a suitable value from the training data, while BoxConstraint (C) equal to 1 means a moderate balance between minimizing errors and maintaining a good margin.
For the CNN-based classifier, the Log-Mel spectrogram is produced using the number of Mel frequency bands set to 64, which represents frequency division to capture sufficient spectral information. The model is trained using the Adam optimizer with 25 epochs, which represents the number of times the model has been trained and is used to allow adequate learning and avoid overfitting. The training progress, shown in **Fig. 4**, shows that the training and validation accuracy follow a similar trend for the proposed model across the iterations, where the convergence indicates learning efficiency and the absence of overfitting. While the validation loss decreases consistently, which assures that the number of epochs (25) is enough to achieve good performance.
Additionally, a mini-batch size of 128 is used, which represents the number of samples used per step used to ensure training is efficient and stable. So, the training parameters are carefully selected to ensure a balance between performance and computational efficiency.
The dataset is separated into training, validation, and testing sets with a ratio of 80%, 10%, and 10%, respectively. This partition ensures an adequate amount of data is used for training the model, while separating the rest of the subsets for validation and testing. The validation set is used during training for parameter tuning and helps in model selection, whereas the testing set is used for evaluating the model on unseen data to estimate the generalization of the model. This ratio provides a balance between learning efficiency and reliable performance evaluation.
The hybrid model combines the outputs of both the SVM and CNN classifiers using a soft voting scheme, where each classifier contributes with a specific weight. Let (w_{SVM}) and (w_{CNN}) denote the weights assigned to the SVM and CNN classifiers, respectively, so that (w_{SVM}) + (w_{CNN}) = 1. Many weight combinations are assessed to study their effect on classification performance as shown in **Table 4** below:

Table 4. The Effect of SVM and CNN Weights on Hybrid Model Accuracy

w_{SVM}	w_{CNN}	Accuracy
0.6	0.4	98.92
0.4	0.6	97.83
0.5	0.5	99.28

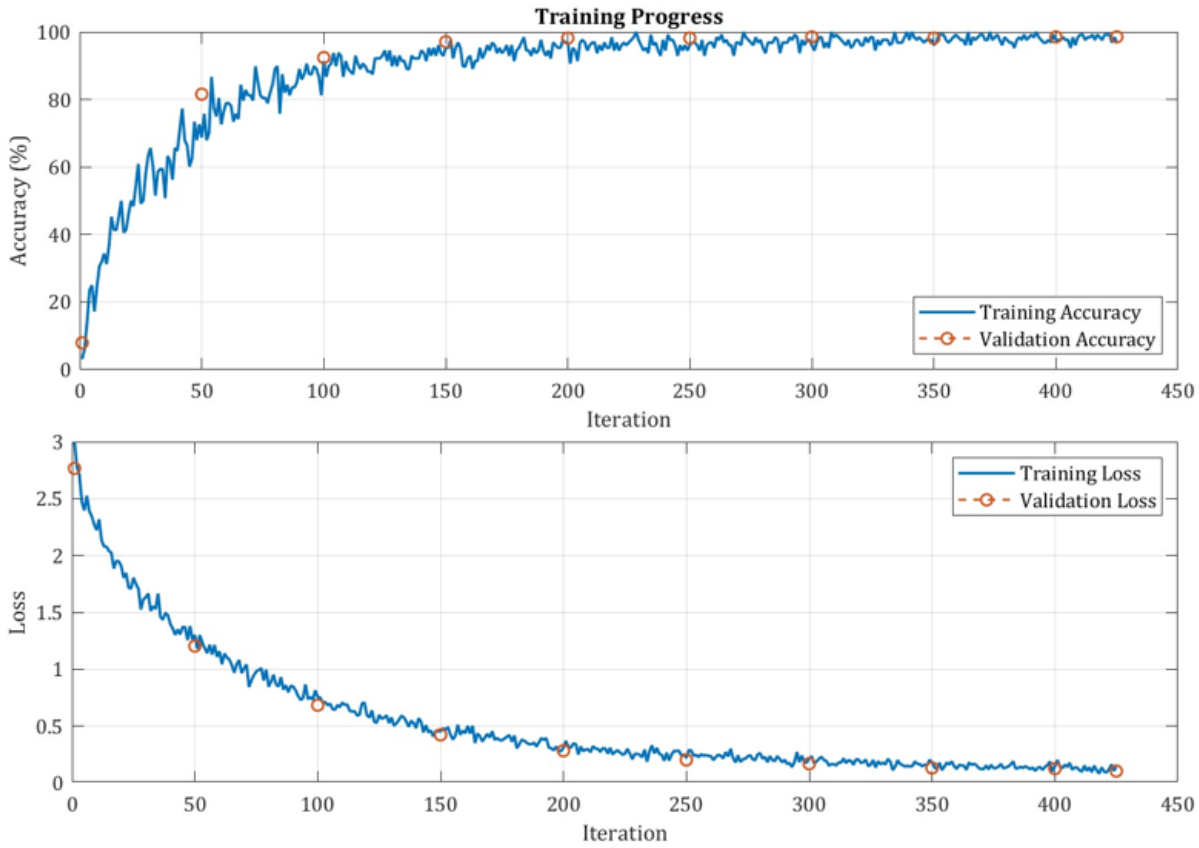

Figure 4. Training Progress of the Model (Accuracy and Loss over Iteration).

The best accuracy is achieved when balanced weights are assigned for both classifiers, equal to 0.5, to achieve optimal performance. In this study, the dataset is created using clean speech signals got from the TIMIT dataset, together with various environmental noise types. The final dataset includes a total of 17 classes, including clean speech. The dataset is arranged into separate folders, where each folder corresponds to a specific sound class. ESC is based on grouping audio signals into different categories. These categories include speech signals, which represent spoken language and babbling speech signals. While the non-speech category is further divided into machine sounds and environmental noise. Machine sounds refer to those generated by mechanical systems and vehicles, such as cars, F-16 aircraft, Factory machinery 1and 2, and Buccaneer 1 and 2 aircraft. In contrast, environmental noise includes ambient and background sounds, such as White noise, pink noise, Hall noise, and Airport noise. This work classifies these different sounds by extracting their features from diverse sound samples, then pre-processing them to train the model to accurately classify them. The sound class labels are automatically given based on the folder names. All audio signals are resampled and segmented into fixed-length samples to ensure a consistent input illustration for both the SVM and CNN-based classifiers, as mentioned above. Several

experiments are conducted by varying the frame length (129, 256, 384, 512) and moment order (50, 60, 70) parameters in order to determine the perfect configuration of the proposed model. The frame length is tested using multiple values, while the moment order is modified to assess its effect on classification performance. For each experiment, the performance of the SVM, CNN, and HCM is evaluated for 5 runs using average classification accuracy, Training and testing times and statistically significant improvements ($p < 0.05$). The results of these experiments are summarized in **Table 5**.

Table 5. Effect of Frame Length and Moment Order on classifiers' performance.

Frame length	Order	Training Time (s) Avg.	Testing Time (s) Avg.	SVM Acc. Avg.	CNN Acc. Avg.	HCM Acc. Avg.	P value <0.05
129	50	246.16812	4.54442	94.078	94.44	96.1	0.005751
129	60	335.42734	4.5338	94.944	94.726	96.39	0.01318
129	70	335.9008	4.51384	96.172	93.428	94.73	0.001442
256	50	331.18524	4.55044	95.952	98.772	98.918	2.98E-05
256	**60**	**327.87446**	**4.64248**	**96.75**	**98.188**	**99.282**	**1.79E-05**
256	70	326.63132	4.57484	96.752	98.124	98.268	0.03278
384	50	330.32866	4.64678	95.234	98.196	98.558	0.0003015
384	60	343.91674	4.78652	94.224	97.254	97.76	0.0009425
384	70	339.97558	4.75858	95.092	97.542	97.758	1.887E-05
512	50	352.16042	4.88608	92.276	97.686	97.83	4.98E-10
512	60	340.18978	4.77072	92.346	96.966	97.616	6.094E-06
512	70	341.59286	4.99628	94.008	97.254	97.76	0.0003264

Table 5 shows the average classification accuracy, the average training time, the average testing time, and statistical significance (P) value results obtained from five independent runs for different combinations of frame length and moment order. Among all tested configurations, a frame length of 256 samples and a moment order of 60 achieved the highest average HCM accuracy of 99.282%, while keeping reasonable computational requirements with an average training time of 327.87 s and testing time of 4.64 s. Furthermore, the obtained p-value (1.79×10^{-5}) is significantly lower than 0.05, emphasize that the performance improvement is statistically significant. For shorter frame lengths (129 samples), the classification accuracy was generally lower, indicating that the extracted segments contained insufficient temporal information. Alternatively, increasing the frame length by more than 256 samples (384 and 512 samples) did not enhance the performance. Likewise, the moment more than 60 orders did not present additional information, which sometimes led to degradation in performance.

From these observations, a frame length of 256 samples and a moment order of 60 are selected as the optimal parameter combination for the proposed ESC framework. The experimental parameters are chosen based on many experiments that are conducted to classify the environmental sound. A frame length of 256 samples was employed, corresponding to approximately 16 ms at a sampling frequency of 16 kHz, which delivers a suitable balance between temporal and spectral resolution. For SVM classifier, a moment order of 60 was selected because it gave the best classification performance among the verified orders. In the hybrid model, equal classifier weights (0.5 for SVM and 0.5 for CNN) were used in the soft-voting stage to ensure balanced contributions from both classifiers. The accuracy result is 99.28%, which indicates both classifiers provide complementary

information, and balanced contributions from the two models lead to the best overall performance.

Generally, the hybrid model often outperforms both the SVM and CNN classifiers across different parameter settings. This indicates that combining handcrafted features with deep learning representations will increase the classifier effectiveness. It was also observed that the SVM classifier is more sensitive at longer frame lengths, where a noticeable drop in performance occurs. In contrast, the CNN model maintains relatively stable performance across different values of both parameters.

4.2 Evaluation Metrics

There are several performance indicators based on the confusion matrix that are used to evaluate the prediction performance of the proposed model. When developing models for binary classification, outcomes are categorized into four types, which are represented in a confusion matrix. The categories are:

- True Positive (**TP**): Successfully estimated positive results.
- True Negative (**TN**): Successfully estimated negative results.
- False Positive (**FP**): Incorrectly classified as positive.
- False Negative (**FN**): Incorrectly classified as negative.

From the confusion matrix, four key metrics are derived to assess the model's effectiveness (Accuracy, Recall, Precision, and F1 score). These metrics are critical for understanding different aspects of the model's accuracy and precision:

- **Accuracy (A)**: is the most widely used indicator for evaluating a system's overall performance; it describes the degree to which a measured value matches the real value. It can be represented by formula (8):

$$\boldsymbol{Accurecy}\ (\boldsymbol{A}) = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{8}$$

- **Precision (P)**: is the proportion of participants correctly identified as positive out of all those identified as positive. It can be computed using formula (9):

$$\boldsymbol{Precision}\ (\boldsymbol{P}) = \frac{TP}{TP+FP} \times 100\% \tag{9}$$

Recall (R): it shows the percentage of all positive samples that are really positive. The formula (10) can be used to compute it:

$$\boldsymbol{Recall}\ (\boldsymbol{R}) = \frac{TP}{TP+FN} \times 100\% \tag{10}$$

F1 Score: is the harmonic mean of precision and recall, providing a balanced measure of classification performance using formula (11).

$$\boldsymbol{F1\ score}\ (\boldsymbol{F}) = 2 \times \frac{P*R}{P+R} \times 100\% \tag{11}$$

The functioning of the SVM, CNN, and HCM models based on these metrics is presented in **Table 6**.

Table 6. Evaluation results of SVM, CNN, and HCM proposed Classifiers.

Classifier	Accuracy	Precision	Recall	F1 Score
SVM	96.03%	96.51%	95.96%	96.23%
CNN	97.83%	98.02%	97.79%	97.91%
HCM	99.28%	99.29%	99.26%	99.28%

The results show that the proposed HCM realizes the highest performance across all evaluation metrics. To further analyze the classification performance, confusion matrices for the SVM, CNN, and HCM are presented in **Figs. 5 to 7**, respectively.

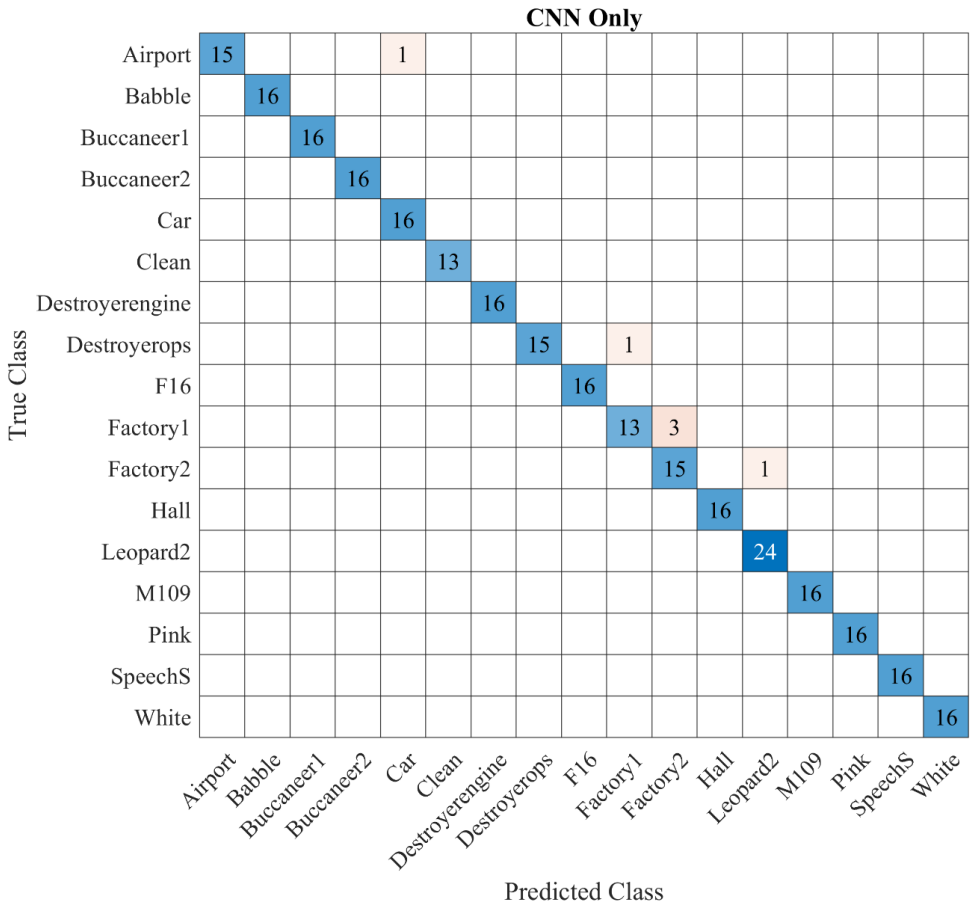

Figure 5. Confusion Matrix for the SVM-Based Classifier.

From **Fig. 5**, it can be observed that the SVM model shows several misclassifications among certain classes, such as confusion between "Destroyerops" and "Babble", "Speech S" and "Babble" as well as between "Factory1" and "Pink". This conduct reflects the sensitivity of the SVM model to feature representation, as it depends on handcrafted features that may not fully capture the complex characteristics of audio signals.

In **Fig. 6**, the CNN model shows improved performance compared to SVM, with fewer misclassifications. However, a small number of samples are incorrectly classified so some confusion still exists, especially in "Factory1" and "Factory2" classes. In contrast, **Fig. 7**, the HCM significantly reduces these misclassifications; most classes are correctly classified, indicating better class separation and more accurate predictions. The misclassification occurs between "Factory 1" and "Factory 2" only.

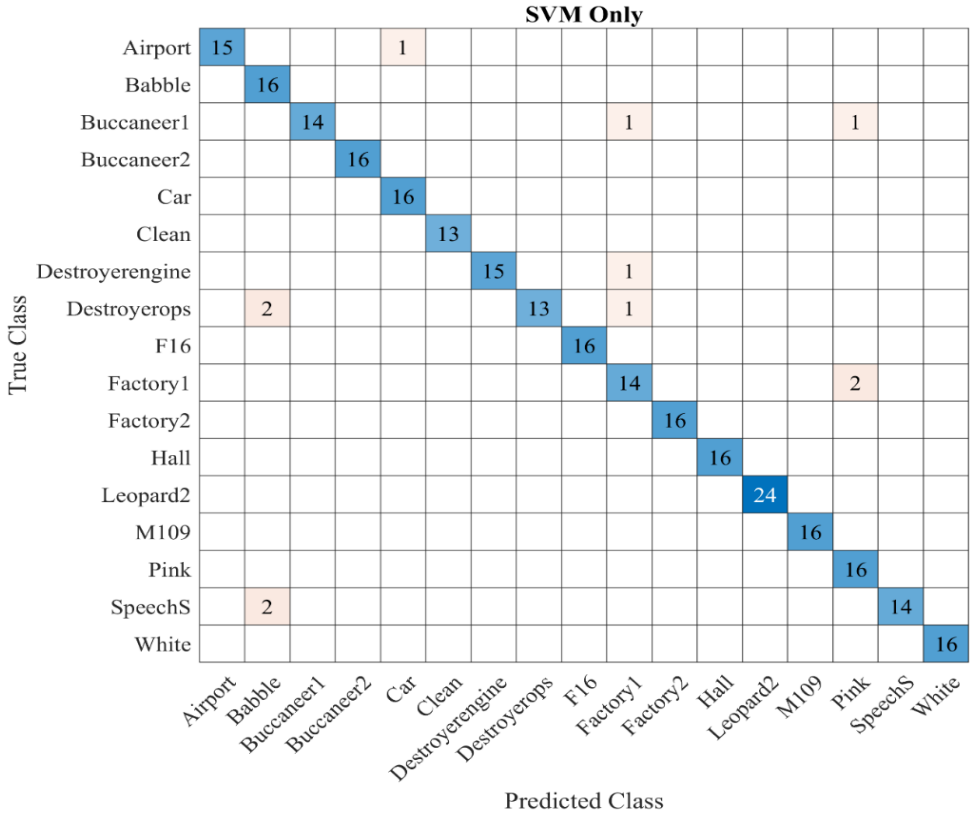

Figure 6. Confusion Matrix for the CNN-Based Classifier.

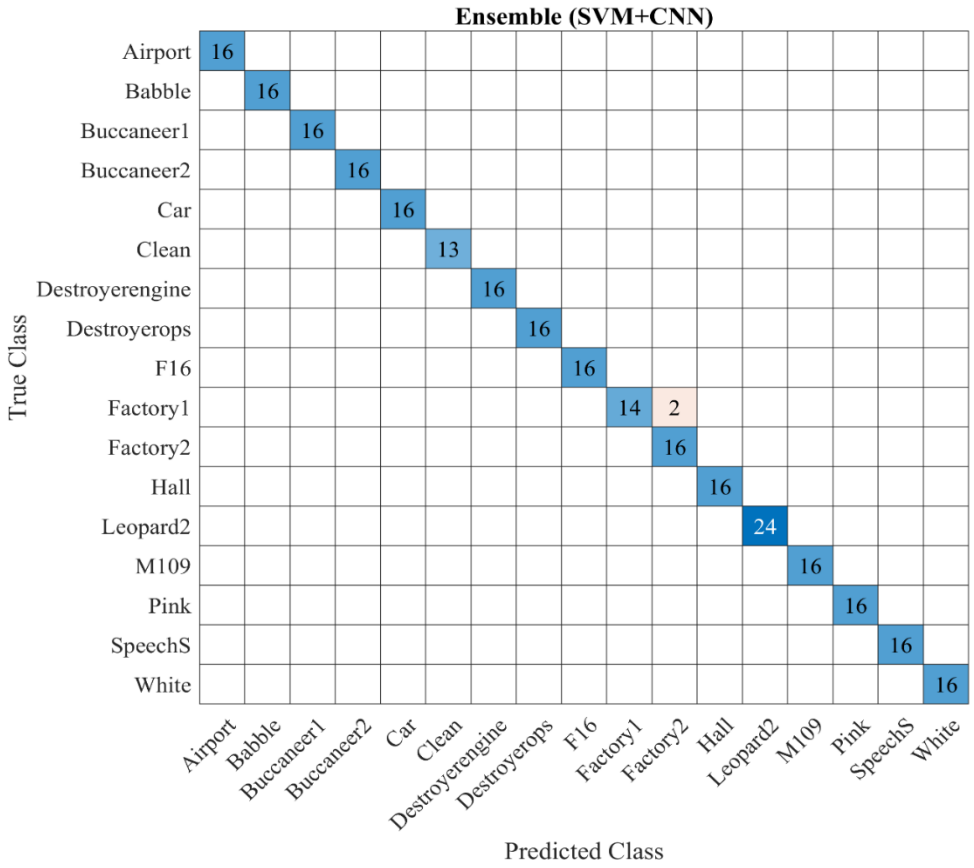

Figure 7. Confusion Matrix for the proposed HCM.

Generally, the proposed hybrid model provides more accurate and reliable classification results, emphasizing its ability to effectively integrate the strengths of both SVM and CNN models.

4.3 Statistical Validation

To further evaluate the generalization capability of the proposed hybrid classifier, five independent experiments (runs) were conducted using different random seeds. In each run, a train-validation-test 80%/10%/10% split was generated, resulting in different data partitions. The classification accuracy obtained in each run is presented in **Table 7.** The proposed hybrid classifier (HCM) achieved a mean accuracy of 98.12% with a standard deviation of 1.29%. Furthermore, the 95% confidence interval was calculated as 98.12 ± 1.13%, corresponding to a confidence range of 96.99%–99.25% as shown in **Table 8**. These results show that the proposed model maintains stable performance across different random data partitions and proves good generalization capability.

Table 7. The classification accuracy for each run

Run	SVM Accuracy %	CNN Accuracy %	(HCM)Accuracy %
1	96.39	95.67	97.11
2	92.78	96.75	96.39
3	96.03	97.11	98.92
4	95.67	98.92	99.28
5	96.75	98.56	98.92

Table 8. Statistical values

Metrics	Value
Mean Accuracy	98.12%
Standard Deviation	1.29%
95% Confidence Interval	98.12 ± 1.13%
Confidence Interval Range	96.99 – 99.25%

To confirm the obtained results in more detail, a one-way Analysis of Variance (ANOVA) test is made on the classification accuracies obtained from five independent runs for each combination of frame length and moment order. This analysis aimed to determine whether the differences noticed in performance among the tested parameter settings are statistically significant. The resulting p-values are described in **Table 5**. All p-values are less than 0.05, demonstrating statistically significant differences among the evaluated configurations. The results verified that the improvements in accuracy are not due to random variations in the data partitioning process. Among the tested configurations, the combination of a frame length of 256 samples and a moment order of 60 achieved the highest average classification accuracy (99.282%), establishing its suitability for the proposed HCM classifier.

4.4 Computational Complexity Analysis

To evaluate computational efficiency, experiments are conducted in a dedicated hardware environment and measured using different assessment metrics. All experiments are conducted using MATLAB 2023a on a workstation equipped with an Intel Core i7-13650HX processor, 24 GB RAM, and a 6 GB GPU. The training and testing times of the hybrid model

were measured on the experimental platform described in **Table 5**. The proposed hybrid model achieved improved classification performance while maintaining an acceptable computational cost.

The results obtained are discussed in the following paragraphs. As shown, this study demonstrated that combining two types of powerful classifiers, such as SVM (Machine learning type) and CNN (Deep learning type), improves the accuracy and performance of the classification process. In this section, the experimental results are discussed and analyzed to evaluate the effectiveness of the proposed hybrid model. The results prove that the hybrid model outperforms the individual SVM and CNN classifiers across all evaluation metrics.

To further evaluate the effectiveness of the proposed hybrid model, a comparison with existing studies is shown in **Table 9**. A related ESC study **(Abdoune et al., 2024; Beritelli et al., 2023)** that used SVM- based machine learning and CNN-based deep learning models, respectively, have been selected for comparison purposes. In addition, other studies such as **(Hu et al., 2024)** that proposed a hybrid CNN-BiLSTM model with attention mechanisms, **(Chen et al., 2025)** which used a MobileNetV2-based Deep Learning Model and **(Chen et al., 2025)** that proposed Spectro MaskNet (Attention-based DL) for environmental sound classification is selected for comparison to highlight the improved performance of deep learning approaches against the existing methods.

Table 9. A comparison between the proposed HCM and the Existing Methods.

Reference	Classifier type	No. of noise	Accuracy
(Abdoune et al., 2024)	SVM (ML)	7	95.19%
(Beritelli et al., 2023)	CNN (DL)	14	96.93%
(Hu et al., 2024)	CNN + BiLSTM (hybrid)	10	98.35%
(Chen et al., 2025)	MobileNetV2-based Deep Learning Model	51	92.16%
(Chen et al., 2025)	SpectroMaskNet (Attention-based DL)	10	97.50%
HCM	SVM +CNN (hybrid)	17	99.28%

This Table summarizes different classification approaches, including machine learning, deep learning, and hybrid models, along with the classifier type, number of noise classes, and the accuracy that was achieved. It can be noticed that traditional machine learning methods, such as SVM, achieve reasonable performance with an accuracy of 95.19% when applied to a limited number of noise classes. Deep learning approaches, such as CNN, show improved performance, achieving 96.93% accuracy. Hybrid models demonstrate superior performance, as shown by the CNN + BiLSTM approach, which achieves 98.35% accuracy.

However, the proposed HCM achieves the highest accuracy of 99.28% while using a larger number of noise classes (17). This improvement proves the effectiveness of combining machine learning and deep learning techniques in one ensemble classifier to improve classification performance.

The SVM model depends on manually extracted features, which are effective in capturing characteristics of the signal in the frequency domain, while the CNN model automatically learns discriminative features from spectrogram representations, capturing complex patterns in both time and frequency domains. By merging these two approaches, the hybrid model is able to reduce the limitations of each classifier. Specifically, misclassifications noticed in the SVM model are reduced by the CNN model, while the overall decision process is represented by the structured feature representation of SVM supports. As a result, the HCM provides more robust and reliable classification performance across different noise types. Finally, the comparison offers a general indication of the effectiveness of the proposed

approach, in spite of different datasets and experimental conditions that are used in the compared studies. Although the proposed hybrid HCM framework demonstrated promising classification performance, several limitations must be acknowledged. First, the experiments were conducted using specific sound signals, which may not fully represent the real-world acoustic environment, as the evaluation considered 17 environmental acoustic categories. Second, the proposed framework is highly computationally complex, requiring extensive processing hardware and enormous amounts of data for training. Furthermore, the proposed approach was evaluated using specific hybrid architecture and did not consider newer deep learning models, such as Transformer-based networks, which might offer better performance.

Future work will focus on evaluating the proposed framework on larger and more diverse datasets, including real-world environmental recordings. Furthermore, advanced architectures such as CNN-Transformer and lightweight deep learning models will be implemented to improve generalization and enhance the capability and classification performance.

5. CONCLUSIONS

This work proposed a hybrid classification framework that combined SVM and CNN for an accurate ESC. The SVM model used a handcrafted orthogonal transform feature, while the CNN model automatically learns deep feature representations from Log-Mel spectrograms. A soft voting strategy is utilized to improve overall decision-making and to get robust and reliable classification decisions. The experiments showed that the proposed model (HCM) outperformed individual classifiers and related existing works in terms of accuracy, precision, recall, and F1-score. In addition, the findings confirmed the efficiency of integrating two powerful approaches, which are ML and DL techniques, with ensemble techniques for handling complex and multi-class audio classification tasks. Overall, for ESC, HCM approach provides a robust and reliable solution and is flexible and scalable solution for real-world applications. Future trends will focus on extending the model to diverse and larger datasets, improving the model's architecture, and exploring advanced feature extraction techniques and other DLs to achieve exceptionally high performance.

Credit Authorship Contribution Statement

Nisreen Talib Abdulhusein: Data Curation, Original Draft Preparation, and Methodology, Manuscript review and editing. Basheera M. Mahmmod: Methodology, Original Draft Preparation, Manuscript review and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abdoune, L., Fezari, M. and Dib, A. 2024. Indoor sound classification with support vector machines: State of the art and experimentation. *International Journal of Computational Methods and Experimental Measurements*, 12(3), pp. 269–279. https://doi.org/10.18280/ijcmem.120307.

Aji, NB., Kurnianingsih, K., Masuyama, N., Nojima, Y. 2024. CNN-LSTM for heartbeat sound classification. *JOIV: International Journal on Informatics Visualization*, 8(2), pp. 735–741. https://dx.doi.org/10.62527/joiv.8.2.2115.

Akter, R., Islam, MR., Debnath, SK., Sarker, PK., Uddin, MK. 2025. A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning. *Digital Signal Processing*, 163, P. 105234. https://doi.org/10.1016/j.dsp.2025.105234.

Albaji, A.O., Rashid, R.B.A. and Abdul Hamid, S.Z. 2023. Investigation on machine learning approaches for environmental noise classifications. *Journal of Electrical and Computer Engineering*, 2023(1), P. 3615137. https://doi.org/10.1155/2023/3615137.

Ali, Y.H., Rashid, R.A. and Hamid, S.Z.A. 2022. A machine learning for environmental noise classification in smart cities. Indonesian *Journal of Electrical Engineering and Computer Science*, 25(3), pp. 1777–1786. https://doi.org/10.11591/ijeecs.v25.i3.pp1777-1786.

Alsouda, Y., Pllana, S. and Kurti, A. 2019. IoT-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest. In *Proceedings of the international conference on omni-layer intelligent systems*, pp. 62–67. https://doi.org/10.1145/3312614.3312631.

Bansal, A. and Garg, N.K. 2022. Environmental sound classification: A descriptive review of the literature. *Intelligent systems with applications*, 16, P. 200115. https://doi.org/10.1016/j.iswa.2022.200115.

Bansal, A. and Garg, N.K. 2023. Environmental sound classification using hybrid ensemble model. *Procedia Computer Science*, 218, pp. 418–428. https://doi.org/10.1016/j.procs.2023.01.024.

Beritelli, L., Borzì, MG., Randieri, C., Avanzato, R., Beritelli, F. 2023. Techniques for recognising and classifying environmental noise using deep learning. In *SYSTEM*, pp. 62–67.

Blanco, V., Japón, A. and Puerto, J. 2022. A mathematical programming approach to SVM-based classification with label noise. *Computers & Industrial Engineering*, 172, P. 108611. https://doi.org/10.1016/j.cie.2022.108611.

Chan, C.-F. and Eric, W.M. 2010. An abnormal sound detection and classification system for surveillance applications. In *2010 18th European Signal Processing Conference*, pp. 1851–1855.

Chen, F., Zhu, Z., C Sun, and Xia, L. 2025. Evaluating metric and contrastive learning in pretrained models for environmental sound classification. *Applied Acoustics*, 232, P. 110593. https://doi.org/10.1016/j.apacoust.2025.110593.

Chen, G., Zhang, B., Ding, Z., Xiao, K., Guan, P., Xiao, X., Wang, X., Yi, H., Hu, H., and Zhang, W. 2025. A lightweight dual branch masking network for environmental sound classification. *Scientific Reports* [Preprint]. https://doi.org/10.1038/s41598-025-33636-w.

Chu, H.-C., Zhang, Y.-L. and Chiang, H.-C. 2023. A CNN sound classification mechanism using data augmentation. *Sensors*, 23(15), P. 6972. https://doi.org/10.3390/s23156972.

Demir, F., Abdullah, D.A. and Sengur, A. 2020. A new deep CNN model for environmental sound classification. *IEEE Access*, 8, pp. 66529–66537. https://doi.org/10.1109/ACCESS.2020.2984903.

Goulão, M., Bandeira, L., Martins, B., and Oliveira, A L. 2024. Training environmental sound classification models for real-world deployment in edge devices. *Discover Applied Sciences*, 6(4), P. 166. https://doi.org/10.1007/s42452-024-05803-7.

Hasan, N.I. 2022. Bird species classification and acoustic features selection based on distributed neural network with two stage windowing of short-term features. *arXiv preprint,* arXiv:2201.00124 [Preprint]. https://doi.org/10.48550/arXiv.2201.00124.

Hasnain, FU., Iqbal, V., Javed, T., and Yasir, M. 2025. Adaptive Boosted Support Vector Machine-random Forest for Environmental Sound Classification. *Journal of Computing & Biomedical Informatics*, 9(02). https://doi.org/10.56979/902/2025.

Hu, R., Hu, K., Wang, L., Guan, Z., Zhou, X., Wang, N., and Ye, L. 2024. Using deep learning to classify environmental sounds in the habitat of western black-crested gibbons. *Diversity*, 16(8), P. 509. https://doi.org/10.3390/d16080509.

Hussein, HA., Hameed, SH., Mahmmod, B.M., Abdulhussain, S.H., Hussain, A.J. 2023. Dual stages of speech enhancement algorithm based on super gaussian speech models, *Journal of Engineering*, 29(09), pp. 1–13. https://doi.org/10.31026/j.eng.2023.09.01.

Ibraheem, Z.H. and Shihab, A.I. 2024. Speech isolation and recognition in crowded noise using a dual-path recurrent neural network. *Iraqi Journal of Science*, pp. 5784–5797. https://doi.org/10.24996/ijs.2024.65.10.37.

Ilahi, A.H.Z., Irwansyah, A. and Oktavianto, H. 2025. Comparative study of CNN architectures for real-time audio-based car accident detection on edge devices. *JOIV: International Journal on Informatics Visualization*, 9(3), pp. 1310–1318. https://dx.doi.org/10.62527/joiv.9.3.2985.

Iqbal, F., Abbasi, A., Almadhor, A., Alsubai, S., and Gregus, M. 2025. Real-time active-learning method for audio-based anomalous event identification and rare events classification for audio events detection. *Frontiers in Computer Science*, 7, P. 1517346. https://doi.org/10.3389/fcomp.2025.1517346.

Islam, M. and Ali, M.N.Y. 2024. Environmental sound classification using feature fusion of mfccs, mel-spectrogram, and chroma. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pp. 3212–3217. https://doi.org/10.1109/ICCIT64611.2024.11021738.

Khayam, S.A. 2003 The discrete cosine transform (DCT): Theory and application. *Michigan State University*, 114(1), P. 31.

Malhotra, A. 2023. Noise identification in urban cities using performance of Random Forest, KNN, and SVM. In *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pp. 1–5. https://doi.org/10.1109/ICSES60034.2023.10465454.

Meedeniya, D., Ariyarathne, I., Bandara, M., Jayasundara, R., and Perera, C. 2023. A survey on deep learning based forest environment sound classification at the edge. *ACM Computing Surveys*, 56(3), pp. 1–36. https://doi.org/10.1145/3618104.

Nasir, R.J. and Abdulmohsin, H.A. 2024. Noise reduction techniques for enhancing speech. *Iraqi Journal of Science*, pp. 5798–5818. https://doi.org/10.24996/ijs.2024.65.10.38.

Nogueira, A.F.R., Oliveira, H.S., Machado, J.J.M., and Tavares, J.M.R.S. 2022. Sound classification and processing of urban environments: A systematic literature review. *Sensors*, 22(22), P. 8608. https://doi.org/10.3390/s22228608.

Piczak, K.J. 2015. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6. https://doi.org/10.1109/MLSP.2015.7324337.

Sankavi, K., Jyothish Lal, G. and Premjith, B. 2024. Deep learning based automatic noisy speech classification for enhanced speech analysis. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, pp. 1166–1171. https://doi.org/10.1109/AIC61668.2024.10730980.

Souadek, R. and Guellil, N., 2025. Environmental sound classification based on STFT features and Convolutional Neural Networks (CNNs). *Przegląd Elektrotechniczny*, 2025(7). https://doi.org/10.15199/48.2025.07.25.

Walden, F., Dasgupta, S., Rahman, M., and Islam, M. 2022. Improving the environmental perception of autonomous vehicles using deep learning-based audio classification. *arXiv preprint arXiv:2209.04075* [Preprint]. https://doi.org/10.48550/arXiv.2209.04075.

Yassen, M.T. and Abdulhussain, S.H. 2025. Discrete cosine transform-driven hybrid orthogonal polynomials: Design and applications in signal processing.

Yousif, S.T., Mahmmod, B.M. 2026. A dual-stage perceptual-harmonic hybrid estimator for speech enhancement. *Journal of Engineering*, 32(3), pp. 173–192. https://doi.org/10.31026/j.eng.2026.03.10.

Zaman, K., Sah, M., Direkoglu, C., and Unoki, M. 2023. A survey of audio classification using deep learning. *IEEE access*, 11, pp. 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015.

Zhang, Y., Huang, Q., Sun, W., Chen, F., Lin, D., and Chen, F. 2024. Research on lung sound classification model based on dual-channel CNN-LSTM algorithm. *Biomedical Signal Processing and Control*, 94, P. 106257. https://doi.org/10.1016/j.bspc.2024.106257.

تصنيف الأصوات البيئية باستخدام نهج هجين SVM-CNN

نسرين طالب عبدالحسين، بشيرة محمدرضا محمود*

قسم هندسة الحاسبات، كلية الهندسة، جامعة بغداد، بغداد، العراق

الخلاصة

تعدّ تصنيف الأصوات البيئية مهمةً بالغة الصعوبة نظرًا لتغير الإشارات الصوتية وتعقيدها. ويتضمن هذا التصنيف تحديد مسار الصوت المرتبط بمختلف الأصوات البيئية. وتُعقّد عدة عوامل هذه المهمة، منها المسافات الطويلة بين المصدر والوجهة، وتداخل الأصوات، وتعدد مصادر الصوت في التسجيلات الصوتية. ولحلّ هذه المشكلة، تقترح هذه الدراسة تقنية تعلّم عميق هجينة لتحسين تصنيف البيانات الصوتية، مع التركيز على خصائص الصوت الأساسية التي أثبتت دقتها في مهام تصنيف الصوت. يدمج إطار التصنيف الهجين المقترح بين آلة المتجهات الداعمة (SVM) والشبكة العصبية الالتفافية (CNN). يستخدم نموذج SVM خصائص العزم المتعامد (DCT) المصممة يدويًا، بينما يتعلم نموذج CNN تلقائيًا التمثيلات من مخططات الطيف Log-Mel. تُجمّع الإشارات المشوّشة استنادًا إلى مجموعة بيانات TIMITو NOISEX-92 لإنشاء إشارات جماعية بناءً على إشارات الكلام النظيفة المدمجة مع أنواع مختلفة من الضوضاء البيئية، مما ينتج عنه 17 فئة من الإشارات المشوّشة، بما في ذلك فئة الإشارة النظيفة. علاوة على ذلك، تُستخدم استراتيجية التصويت المرن لتحسين عملية اتخاذ القرار بشكل عام. يحسب هذا النموذج متوسط الاحتمالية (أو المتوسط المرجح) لكل فئة عبر جميع النماذج، ثم يختار الفئة ذات أعلى متوسط احتمالية كناتج نهائي، مما يؤدي إلى تصنيف أكثر دقة وموثوقية. تُثبت النتائج التجريبية أن النموذج الهجين المقترح يحقق أداءً أفضل مقارنةً بالطرق الحالية والمصنفات الفردية، حيث تصل دقته إلى 99.28%، بالإضافة إلى تحسينات في الدقة والاستدعاء ومقياس F1. تعكس هذه النتائج كفاءة دمج أساليب التعلم الآلي والتعلم العميق لتصنيف الأصوات البيئية.

الكلمات المفتاحية: الشبكة العصبية التلافيفية، آلة المتجهات الداعمة، التعلم العميق، التعلم الآلي، التصويت الناعم تصنيف الاصوات البيئية.