Intelligent System for Automatic Recognition of Environmental Sounds Using Optimal Feature Fusion and Ensemble Deep Learning Technique

Divya Lakshmi. S [*1,2,] N. Suresh Kumar [3]
[1]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
[2]Department of Computer Applications, Marian College Kuttikkanam Autonomous,
Kerala, India
[3]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
E-mail: divyabalu19@gmail.com, sureshkumar@ klu.ac.in, divyabalu19@gmail.com
[*]Corresponding author

Keywords: automatic recognition, environmental sounds, ESC, feature fusion, ensemble learning

Received: April 5, 2025

Environmental sound classification (ESC) is a challenging task due to the unstructured and overlapping nature of ambient sounds, which differ significantly from speech and music. Problems such as class imbalance, limited labeled samples, and high inter-class similarity hinder the performance of traditional classifiers. In this study, we propose a robust ESC system that combines optimal spectrum feature fusion with a stacked ensemble learning strategy. Specifically, we extract three types of spectral features—log Mel spectrum, log–log Mel spectrum, and Mel spectrograms—from environmental audio signals using the DenseNet-161 architecture. These features are then optimally fused using the Boosted Reptile Squirrel Search (BRSS) algorithm to capture both fine- and coarse-grained frequency patterns. For classification, we employ a two-level ensemble model: four classical machine learning classifiers (Linear Regression, Decision Tree, Random Forest, and Support Vector Machine) in the first stage, followed by a Bayesian Tensorized Neural Network (BTNN) for final prediction. Experimental results on three benchmark datasets—ESC-10, ESC-50, and UrbanSound8K—demonstrate that our fused spectrum feature approach achieves an accuracy of 98.98%, surpassing individual feature types and outperforming state-of-the-art models such as Convolutional Recurrent Neural Network (CRNN), EnvNet, and DualResNet. These results highlight the effectiveness and superiority of our proposed method for environmental sound classification.

Povzetek: Predlagan ansambelski pristop z optimalno fuzijo spektralnih značilk doseže 98,98 % natančnost in preseže obstoječe modele za klasifikacijo okoljskih zvokov.

1 Introduction

Modern technology relies on environmental sound classification (ESC) to automatically identify and classify environmental noises. ESC includes animal calls, traffic noise, weather patterns, and industrial machinery, unlike standard sound recognition systems that focus on speech or music [1]. This field has garnered interest for its possible uses in environmental monitoring, urban acoustic analysis, wildlife conservation, and smart city development [2]. ESC techniques are robust enough to discriminate sound classes in complicated and variable real-world situations [3]. Additionally, the variety of sound sources and lack of standardized databases hinder researchers in this subject [4]. Even so, recent advances in machine learning, especially deep learning, have improved ESC system accuracy and scalability. Hybrid designs improve sound recognition performance and versatility by combining multiple models [5]. These designs handle sound recognition's many issues by combining deep learning, classical machine learning, and signal processing. Recurrent neural network (RNN) and convolutional neural network (CNN) are often used in hybrid architectures to capture spatial features from spectrograms or Mel-frequency cepstral coefficients (MFCCs) and model temporal events [6]. A multi-stage procedure extracts significant features from raw audio signals before feeding them into the network for classification in hybrid sound recognition architecture [7]. Time-frequency alterations like short-time Fourier transform (STFT) or wavelet transform can convert audio inputs for neural network analysis [8]. Hybrid sound recognition architectures can also use ensemble approaches, where numerous models are trained separately and their predictions are integrated. By lowering variance and bias, ensemble approaches like bagging and boosting improve system robustness and generalization [9]. ML/DL techniques are essential to sophisticated environmental sound detection systems [10].

With the rise of digital audio data, sound recognition systems cannot handle the complexity and variety of real-world ambient noises. DL algorithms learn hierarchical data representations from raw audio waveforms, giving a powerful solution [11]. DL algorithms, especially CNNs and RNNs, have shown great promise in automatically learning discriminative features from audio data for environmental sound detection [12]. CNNs are good at capturing spatial patterns in spectrograms or other time-frequency audio representations, while RNNs enjoy modeling temporal relationships in sequential audio data. CNNs and RNNs complement each other; therefore, researchers can use a hybrid framework to improve ambient sound identification performance. ML algorithms that understand patterns and correlations from labeled training data are essential for ambient sound detection [13]. The k-nearest neighbors (k-NN), random forest (RF) and support vector machine (SVM) are used for classification tasks to map audio signal properties to predetermined sound categories. Unsupervised and semi-supervised learning methods allow the investigation of latent structures in ambient sound samples. Deep learning and machine learning can improve automatic environmental sound recognition, enabling innovative applications in wildlife monitoring, urban sounds cape analysis, and healthcare [14]. Hybrid architecture in sound detection helps several sectors autonomously identify and analyze environmental sounds. Hybrid sound recognition systems can identify and classify animal vocalizations, weather patterns, and ecological problems in environmental monitoring [15]. Besides environmental monitoring, hybrid sound recognition is used in smart cities, industrial automation, and healthcare. Smart cities can use sound recognition technology to monitor traffic, detect emergency sirens, and spot unusual events like accidents and disturbances [16]. Hybrid architecture in sound recognition can improve automation, efficiency, and decision-making across industries [17]. These hybrid systems use deep learning and machine learning to adapt to varied contexts and learn from enormous amounts of audio data to increase accuracy and robustness. As they enable proactive environmental risk management, early anomaly detection, and quick response to catastrophic events, such systems can save money, optimize resources, and improve quality of life [18]. Environmental noises' high frequency, loudness, and duration unpredictability presents a problem [19]. It makes sound classification and acoustic pattern differentiation challenging, especially in loud or dynamic contexts. Hardware limits, data privacy concerns, and interoperability issues may complicate hybrid sound recognition system adoption, requiring careful planning and mitigation. Future hybrid sound recognition research could improve system scalability, adaptability, and real-time performance to satisfy changing application needs [20].

Major contributions

The key contributions of this work are summarized as follows:

- Hybrid Deep Ensemble Architecture: We propose a stacked ensemble learning framework that integrates four classical machine learning classifiers—Linear Regression, Decision Tree, Random Forest, and Support Vector Machine—with a Bayesian Tensorized Neural Network (BTNN). This hybrid architecture enhances classification precision and generalization compared to standalone classifiers and deep models.
- Adaptive Feature Fusion Strategy: We introduce a meta-heuristic optimization approach using the Boosted Reptile Squirrel Search (BRSS) algorithm to fuse multiple spectrum-based representations, including log Mel, log–log Mel, and Mel spectrogram features. This fusion strategy improves feature discriminability and mitigates overfitting.
- End-to-End Recognition Pipeline: We develop a novel end-to-end ESC system that combines DenseNet-161-based deep feature extraction, adaptive feature fusion, and two-level ensemble learning, offering both high accuracy and computational efficiency suitable for real-world deployment.
- Advancement Over State-of-the-Art: Evaluated on benchmark datasets ESC-10, ESC-50, and UrbanSound8K, the proposed system achieves a top accuracy of 98.98% using fused features, significantly outperforming prior state-of-the-art methods such as CRNN, EnvNet, and DualResNet.

2 Review of literature

In this section, we provide an overview of the literature concerning the recognition of environmental sounds using ML and DL techniques. Table 1 presents a summary of the research gaps identified in existing state-of-the-art works on environmental sound recognition.

2.1 State of art works

Yildirim et al. 2024 [21] suggested a hybrid model for PD detection using sound data. Sound input is converted into spectrograms, and three CNN architectures extract unique feature maps. The arithmetic optimization algorithm (AOA), an innovative metaheuristic optimum method, helps fuse and choose these varied feature maps. SVM and KNN classifiers are then used for classification. With an accuracy rating of 98.19%, the suggested model diagnoses PD well. The suggested model is also compared to Mel-frequency cestrum coefficients feature maps. RF classifier achieved the highest accuracy of 93.98%.

Mekruksavanich et al. 2023 [22] have explored the field of DL for humanoid movement acknowledgment and puts up successful methods for recognition. In order to find the best architecture for activity recognition, the study first investigates various convolutional neural networks. A channel attention mechanism–integrated hybrid convolutional neural network is the end result of further efforts. The network is able to effectively detect different human movements in daily life because to this technique, which allows it to hierarchically discriminate deep spatio-temporal properties. The model outperforms methods and effectiveness in improving recognition accuracy with 98.92%, 98.80%, and 98.45% accuracy rates, respectively.

Ansari et al. 2023 [23] provided a new architecture for three-way neural networks that can model speech sequences with direct context-awareness: transformer, prior trained dual-path recurring neural network, and transfer learning. Investigational outcomes show that the suggested model outperforms seven advanced deep learning-related architectures on a variety of objective criteria. It outperforms its closest competition and proves its speech separation efficacy with usual development of 4.60% in brief goal comprehension, 14.84% in from source to distort ratio, and 9.87% in scale-invariant proportion of noise to signal.

Wang et al. 2023 [24] came up with a new deep learning strategy for multi-class classification, which includes ternary and binary tasks, by merging a CNN with a LSTM system. This CNN-LSTM hybrid outperforms both conventional ML and ultramodern DL models in ternary classification. By streamlining the process and doing away with manual processes, the suggested method provides a more effective diagnostic tool for doctors, which could make neurologists' jobs easier when it comes to diagnosing epilepsy. It has been widely accepted in acoustic signal processing area that the frequency band has more characteristics information about target sound than the time series.

Rashmi et al. 2023 [25] exploring the use of CNN to mechanically study topographies from audio signals of the English alphabet. They used MFCC-based features and the other that uses a hybrid feature extraction method including LM, MFCC, chroma, spectral contrast, and Tonnetz features. The suggested strategy outperforms current CNN methods using single extraction of features techniques in terms of taxonomic accuracy, and it does this by combining multiple sets of features and training them using separate CNNs. Results show that CNNs work well for sound recognition, especially when combined with hybrid feature extraction techniques; this opens up exciting new possibilities for research in the area.

Jahangir et al. 2023 [26] discussed the neural network system that because babies primarily use crying to express what they need; parents must be extremely careful and keep a close eye on them at all times. With recall, f1-score, and precision rates of 98.39%, 98.05%, and 98.72%, respectively, stacked classifier CNN-SCNet stood up as the most successful. An encouraging answer for worried parents, this study highlights the importance of strong ML models like CNN-SCNet in improving the capacity of baby monitoring systems to identify screams in busy home settings.

Ullo et al. 2020 [27] have proposed the hybrid ESC model based on OAS to extract meaningful samples from each sound class. The time-frequency-amplitude representation is generated by subjecting these representative samples to short-Time Fourier Transform (STFT). These features were trained using prior training AlexNet and VGG-16 networks. Tests on the ESC-10 dataset show that the suggested strategy is as good as, or better than, current state-of-the-art approaches, with accuracies ranging from 87.9% to 95.8%.

Liu et al. 2023 [28] have suggested a ship-radiated extremely fine noise detection scheme consuming amplitude–frequency–time domain multi-scale characteristics and an adaptive generalized network. Superior signal decomposition methods like permutation entropy-based analysis generate six learnable amplitude–time–frequency components from ship-radiated noise signals. 1D CNN and LSTM systems integrate aggregated seasonal characteristics and excellent regional data to focus on time–frequency information in MFAGNet. Testfindings show that MFAGNet outperforms baseline approaches in distinguishing 12 ship noises from ShipsEar dataset and classifying four common ship types from multiple datasets with 98.89% accuracy.

Chen et al. 2024 [29] have promoted heterogeneous coding techniques for comprehensive SNN architecture design. They present a hybrid neural coding and learning system that combines many neuroscience-discovered neural coding schemes. The system also includes unique layer-wise learning algorithms for hybrid coding SNNs and a variable neural coding allocation strategy for task-specific needs. The experiments on image categorization and localization of sounds show that the planned outline outclasses advanced SNNs in accuracy, inference delay, energy consumption, and noise robustness. This study illuminate's hybrid neural coding architectures, paving the way for high-performance neuromorphic devices.

Demir et al. 2020 [30] employed a CNN model trained end-to-end with spectrogram data to improve classification accuracy through the inclusion of deep features. In order to construct a feature vector, the fully linked layers of the suggested CNN model are used to extract deep features. To measure its efficacy, the K-NN ensemble classifier takes this vector as input. The proposed CNN-based technique is effective in ESC tasks, as demonstrated by the remarkable classification accuracies of 96.23% and 86.70% on the DCASE-2017 ASC and UrbanSound8K datasets, respectively.

2.2 Problem description and definition

The task of environmental sound classification (ESC) presents several core challenges that hinder the development of robust, scalable, and accurate models. These challenges are particularly evident in real-world audio environments, where sounds are diverse, overlapping, and often unstructured [31]. The key issues include:

- Unstructured Nature of Environmental Sounds: Unlike speech or music, environmental sounds lack consistent temporal and spectral patterns, making them more difficult to model and classify effectively.
- Data Imbalance and Limited Samples: Many ESC datasets contain an uneven distribution of sound classes and limited recordings per class, which can bias model performance and reduce generalization.
- Large Number of Sound Categories: ESC often involves classification across a wide range of classes with overlapping acoustic characteristics, increasing the complexity of the task.
- Computationally Intensive Feature Extraction: Extracting high-quality features from raw audio—especially using complex or non-linear methods—requires substantial computational resources, making it difficult to scale or deploy in resource-constrained environments.
- Risk of Overfitting: Deep learning models trained on limited or imbalanced data are prone to overfitting, resulting in poor generalization on unseen data.
- Lack of Robustness in Existing Models: Many existing ESC models show inconsistent performance across datasets and real-world scenarios due to suboptimal architectures or feature representation limitations.

These challenges highlight the need for novel, efficient, and generalizable approaches to environmental sound classification that can improve both accuracy and computational.

Table 1: Summary of research gap

Ref.	Feature fusion	Classifier	Findings	Research gaps
[21]	AOA-CNN	k-NN and SVM	Accuracy 98.19%	They use only one single vector to extract features
[22]	UNet	Hybrid CNN	Accuracy 98.92%	It is challenging to identify sounds from limited samples
[23]	DenseNet	SVM and RF	Accuracy 85.965%	Difficult to achieve the complex time–frequency features
[24]	Space–time algorithm	CNN-LSTM	Accuracy 91.253%	Fail to consider the temporal structure, frequency characteristics
[25]	LM, MFCC, and CST	CNN	Accuracy 87.523%	Insufficient structural information of the audio signal
[26]	SCNet	CNN	Precision 98.72%	Limited by number of samples, the network cannot learn more features
[27]	OAS, STFT, VGG-16	k-NN and SVM	Accuracy 95.8%	Often demand computationally intensive operations
[28]	MFAGNet	LSTM	Accuracy 98.9%	Time-consuming and high-effort task
[29]	UNet and DenseNet	Spiking neural networks (SNN)	Accuracy 90.56%	Achieving the classification effect of feature is complex issue
[30]	Deep CNN	k-NN and SVM	Accuracy 96.23%	The feature fusion is not effectively handled which limits the performance

3 Methodology

Our proposed system addresses the challenge of environmental sound classification by transforming raw audio signals into image-based representations and leveraging a stacked ensemble deep learning (DL) architecture for robust and accurate recognition. In Figure 1 the flowchart illustrates the complete methodology of the proposed ESC system. Raw audio signals are first collected from benchmark datasets and undergo preprocessing and augmentation. These signals are then converted into spectrograms using log-Mel, log–log-Mel, and Mel transformations. The extracted features are fused optimally using the Boosted Reptile Squirrel Search (BRSS) algorithm. The fused features are passed to a stacked ensemble learning model consisting of a first layer of machine learning classifiers (Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine), followed by a Bayesian Tensorized Neural Network (BTNN) as the meta-learner. The final output is the predicted environmental sound class.

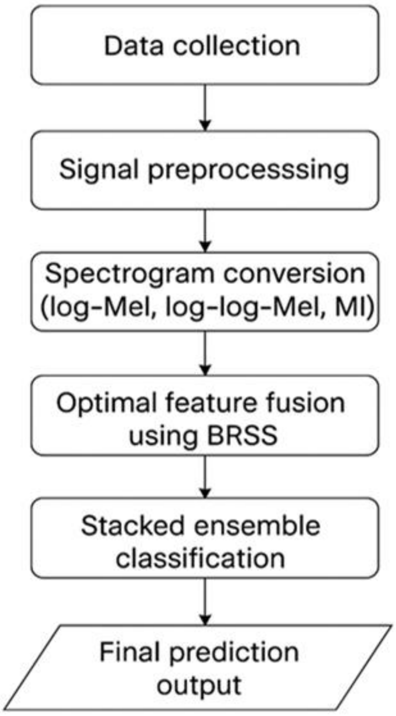
Data collection
Signal preprocesssing
Spectrogram conversion (log-Mel, log-log-Mel, MI)
Optimal feature fusion using BRSS
Stacked ensemble classification
Final prediction output

Figure1: Workflow of the proposed environmental sound classification system

3.1 Overview of the proposed system

As shown in Figure 2, the system comprises five main steps. First, raw environmental sound signals undergo preprocessing and augmentation to improve quality and enhance variability. Second, the preprocessed signals are transformed into spectrogram representations using log-Mel based methods, which convert the audio into a two-dimensional time-frequency format. Third, these spectrograms are input into DenseNet-161, a convolutional neural network that extracts high-level discriminative features. Fourth, the extracted deep features are fused using the Boosted Reptile Squirrel Search (BRSS) algorithm to maximize feature diversity and reduce overfitting. Finally, in the fifth stage, the fused features are classified using a stacked ensemble learning approach. This ensemble includes four base classifiers—linear regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM)—whose outputs are combined by a meta-learner, the Bayesian Tensorized Neural Network (BTNN), to produce the final prediction.

3.2. Stacked ensemble deep learning architecture

The classification stage utilizes a two-layer stacked ensemble DL architecture:

- First Layer (Base Learners): Four traditional machine learning classifiers—Linear Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—are trained independently on the optimized fused features. Each base learner outputs class probabilities for each input sample.

- Second Layer (Meta Learner): The outputs from the first-layer classifiers are concatenated and input into a Bayesian Tensorized Neural Network (BTNN). BTNN leverages tensor-train decomposition and Bayesian inference to capture high-dimensional interdependencies, enhancing classification robustness and effectively modeling uncertainty.

3.3 Feature extraction using DenseNet-161

Audio spectrograms are processed by a pre-trained DenseNet-161 model, which efficiently extracts deep frequency domain features. DenseNet's dense connectivity and feature reuse capabilities enable it to capture subtle variations and complex patterns in environmental sound data, resulting in richer and more informative feature representations.

3.4 Optimal feature fusion with BRSS

To further enhance model generalization, the Boosted Reptile Squirrel Search (BRSS) algorithm is applied for feature fusion. BRSS, inspired by animal foraging behavior, optimally combines feature vectors extracted from DenseNet-161 across multiple spectrogram types.

This approach enhances the discriminative power of the final feature set while mitigating the risk of overfitting. Technical equations and algorithmic steps for BRSS are provided in Section 5.1. To enhance the representational power of spectral features and overcome the limitations of traditional fusion approaches, we adopt an optimization-driven feature fusion strategy based on the Boosted Reptile Squirrel Search (BRSS) algorithm. While conventional fusion techniques such as Principal Component Analysis (PCA), direct concatenation, and Deep Canonical Correlation Analysis (DCCA) have been widely used for combining audio features, they often suffer from redundancy, suboptimal weighting, or lack of adaptability to nonlinear feature interactions [32]. Unlike these methods, our BRSS-based approach leverages metaheuristic search principles to dynamically select and combine complementary features, thereby minimizing redundancy and maximizing classification-relevant information. Metaheuristic optimization has proven especially effective in complex search spaces where gradient-based or fixed-rule strategies fail to generalize [33]. The integration of BRSS enables adaptive feature weighting and selection tailored to the target dataset, offering a significant performance boost over both shallow and deep baseline fusion models.

3.5. Final classification using BTNN

The Bayesian Tensorized Neural Network (BTNN) in the meta-learner layer efficiently handles high-dimensional fused features via tensor-train decomposition. Its Bayesian modeling framework incorporates uncertainty, leading to improved accuracy and reliability in environmental sound classification. The detailed training process for BTNN is outlined in Algorithm 2.

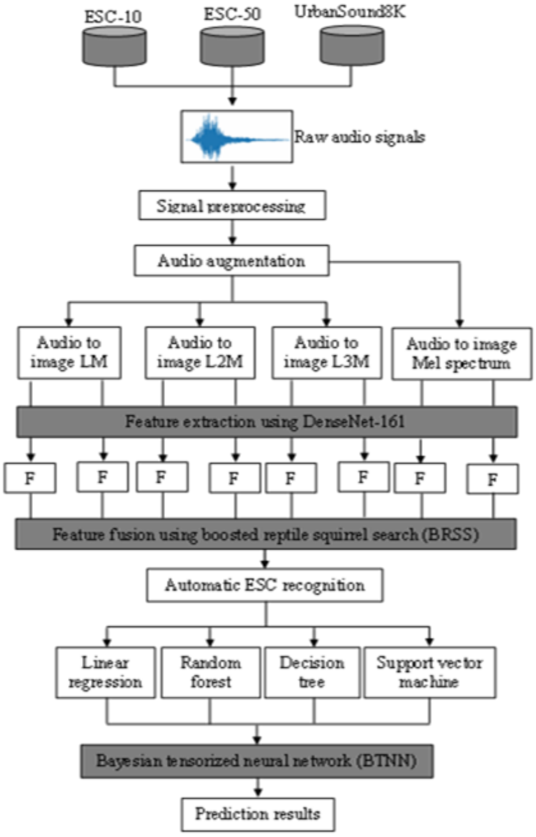

Figure 2. System architecture for automatic ESC using BRSS-based feature fusion and stacked ensemble DL with BTNN meta-learner

4 Background

This section provides brief descriptions of the machine learning algorithms used as base learners in our stacked ensemble model. These classical classifiers serve as the foundation for the first-layer predictions in our architecture and are chosen for their complementary strengths in handling diverse feature patterns in environmental sound classification.

4.1 Linear regression (LR)

Linear Regression models the relationship between input features and the target variable as a linear function:

$$\hat{y} = \mathrm{Xw} + \mathrm{b} \qquad (1)$$

where X is the feature matrix, www is the weight vector, and bbb is the bias term. Model parameters are estimated by minimizing the mean squared error (MSE) between predicted and actual values [34].

4.2 Decision Tree (DT)

A Decision Tree recursively partitions the feature space by selecting optimal feature splits at each node to maximize information gain or minimize impurity (e.g., Gini index), resulting in a hierarchical structure for classification.

4.3 Random Forest (RF)

Random Forest is an ensemble of decision trees trained on different bootstrapped subsets of the data. It uses majority voting among trees to make robust predictions and helps reduce variance and overfitting.

4.4 Support Vector Machine (SVM)

SVM finds the optimal hyperplane that maximizes the margin between classes. When the data is non-linearly separable, kernel functions are employed to project it into a higher-dimensional space for improved class separation.

5 Algorithmic details

This section describes the core algorithms used in the proposed system: the BRSS algorithm for optimal feature fusion and the BTNN classifier for final prediction.

5.1 Boosted Reptile Squirrel Search (BRSS) for feature fusion

BRSS is a metaheuristic optimization algorithm inspired by reptilian and foraging behaviors, designed to fuse feature vectors extracted by DenseNet-161. It updates candidate solutions iteratively to improve classification performance while mitigating overfitting [35].

$$\mathrm{F}S_{u,h} = \mathrm{F}S_k + \mathrm{U}(0,1) \times (FS_i - FS_l)(u = 1,2,\dots,m)(h = 1,2,\dots,f) \quad (2)$$

where $FS_{u,h}$ represents the u-th position of the squirrel in the h-th dimension. We calculate the fitness value corresponding to each squirrel position as follows.

$$\mathrm{F}S_{u,h} = \mathrm{F}S_k + \mathrm{U}(0,1)(FS_i - FS_l)(u1,2,\dots,m)(h1,2,\dots,f) \quad (3)$$

where y addresses the ongoing emphasis, the e1 represents an irregular number in the scope of [0, 1] and the H_v is a skimming consistent. $e_1 \geq o_{fo}$ addresses the likelihood of the presence of hunters. The dg is the arbitrary sliding distance steady. At the point when $e_1 \geq o_{fo}$ without any hunters in the woodland, squirrels skim to find food squirrels have free rummaging exercises. As indicated by the occasional consistent and occasional recognition condition (A_v^r) are determined to decide if entering winter.

$$A_v^r = \sqrt{\sum_{l=1}^{f}\left(FS_{s,l}^r - FS_{g,l}\right)^2} \quad (4)$$

where A_v^r addresses the element of the issue. $\left(FS_{s,l}^r - FS_{g,l}\right)$, separately, signify the squirrel on the walnut tree (best arrangement) and the squirrels on oak trees.

$$A\frac{10e^{-6}}{(365)^{y/(y_m/2.5)}}_{min} \quad (5)$$

where y_m represents the maximum number of iterations. When $FS_{my}^{new} = FS_k$ the positions of those flying squirrels without food sources are updated as follows

$$\mathrm{F}S_{my}^{new} = \mathrm{F}S_k + \mathrm{levy}(m) \times (FS_i - FS_k) \quad (6)$$

Levy Flight allows squirrels to find new locations close to their current sweet spot by

$$levy = 0.01 \times \frac{e_s \times \sigma}{|e_v|^{\frac{1}{\beta}}} \quad (7)$$

where, ra and rb are two typically circulated arbitrary numbers in the scope of [0, 1]. The β is an example boundary of the Duty appropriation, utilized to describe the circulation's shape. The σ is a boundary inside the Duty flight model, overseeing the extent of step lengths. It imitates the jump distance of a Duty flight, processed as follows.

$$\sigma = \left(\frac{\Gamma(1+\beta) \times sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}}\right)^{1/\beta} \quad (8)$$

where $\Gamma(z) = (z-1)$. By using the BRSS algorithm for feature fusion, we effectively combine the deep features extracted from DenseNet-161 with other relevant features, resulting in an optimized feature set that enhances the discriminative power of the classification model while mitigating the risk of overfitting. Algorithm 1 describes the working process of optimal feature fusion using BRSS.

Algorithm 1: Optimal feature fusion using BRSS

```
Input   : Optimization problem information
Output :  Feature fusion
Set control parameters population (m),
Generate random locations for n flying squirrels using
$FS_{u,h} = FS_k + U(0,1)\times(FS_i - FS_l)(u = 1,2,...,m)(h = 1,2,..., f)$
Evaluate the fitness of each flying squirrel's location.
Sort flying squirrel locations by fitness value.
The best value is defined as the squirrel on the pecan tree
While do
For t = 1 to n
Update flying squirrel locations which are on oak trees and moving towards pecan trees using
Equation $FS_m^{r+1} = \begin{cases} FS_m^r + f_h \times H_v \times (FS_g^r - FS_m^r), e_1 \geq o_{fo} \\ Randomlocation \qquad otherwise \end{cases}$
Update flying squirrel locations which are on normal trees and moving towards oak trees
using $FS_m^{r+1} = \begin{cases} FS_m^r + f_h \times H_v \times (FS_g^r - FS_m^r), e_1 \geq o_{fo} \\ Randomlocation \qquad otherwise \end{cases}$
Evaluate the fitness of each flying squirrel's location.
Update flying squirrel locations which are on normal trees and moving towards pecan trees
using $FS_m^{r+1} = \begin{cases} FS_m^r + f_h \times H_v \times (FS_g^r - FS_m^r), e_1 \geq o_{fo} \\ Randomlocation \qquad otherwise \end{cases}$
Evaluate the fitness of each flying squirrel location.
End
Calculate seasonal constant $A_v^r = \sqrt{\sum_{l=1}^{f} (FS_{s,l}^r - FS_{g,l})^2}$
Update the minimum value of seasonal constant ( $A_{\min}$ ) $A_{\min} = \frac{10e^{-6}}{(365)^{y/(y_m/2.5)}}$
If (Seasonal monitoring condition is satisfied)
Randomly relocate flying squirrels on normal trees using
$FS_{my}^{new} = FS_k + levy(m)\times(FS_i - FS_k)$
Evaluate the fitness of each flying squirrel's location
End
The best value is defined as the squirrel on the pecan tree
End
The location of the squirrel on the pecan tree is the final optimal solution
End
```

5.2 Bayesian Tensorized Neural Network (BTNN) for classification

The Bayesian Tensorized Neural Network (BTNN) serves as the meta-learner in the second layer of our stacked ensemble architecture. It is designed to model high-dimensional fused feature representations generated by the Boosted Reptile Squirrel Search (BRSS) algorithm, using a compact tensor-train (TT) structure and Bayesian inference. This design enables efficient parameterization while capturing uncertainty in model predictions, improving generalization performance on unseen environmental sound samples.

The BTNN models the high-dimensional weight tensor using tensor-train decomposition. A tensor of order is represented as a sequence of 3-way tensor cores:

$$\mathrm{M} = \{G^{(1)}, G^{(2)}, \dots, G^{(d)}\} \qquad (8)$$

Each core tensor $G^{(i)} \in \mathbb{R}$^ ($r_{i-1} \times n_i \times r_i$), where r_i are the TT-ranks, and n_i is the size of the i-th dimension of the input.

Given a matrix Z, the tensor-train factorization is defined as:

$$\mathrm{Z} = G^{(1)} \times_1 G^{(2)} \times_2 \cdots \times_{d-1} G^{(d)} \qquad (9)$$

This decomposition allows for a compact representation of the weight matrix using significantly fewer parameters [37].

A single-layer BTNN prediction can be expressed as:

$$\hat{y} = \mathrm{f}(W \cdot x + b) \qquad (10)$$

where W is the TT-decomposed weight tensor, x is the fused feature input, b is the bias, and $f(\cdot)$ is the activation function (e.g., ReLU or SoftMax).

To incorporate uncertainty and avoid overfitting, BTNN is trained using a Bayesian objective that combines the standard cross-entropy loss with a regularization term based on the Kullback-Leibler (KL) divergence between the approximate where $q(\theta)$ denotes the posterior distribution and $p(\theta)$ represents the prior distribution over the model parameters θ.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y_i}) + \lambda \cdot K\mathrm{L}\big(q(\theta)\big|p(\theta)\big) \quad (11)$$

where λ is the regularization coefficient, and θ encompasses all TT-cores and biases in the model.

The posterior distribution over parameters is estimated as:

$$\mathrm{q}(\theta) \propto \exp\left(-\frac{1}{T}\mathcal{L}(\theta)\right) \qquad (12)$$

Where T is a temperature scaling factor that controls the sharpness of the posterior distribution.

By integrating tensor-train decomposition and Bayesian inference, the BTNN provides an expressive yet computationally efficient approach for environmental sound classification. It serves as a powerful meta-learner that consolidates predictions from the base learners in the stacked ensemble architecture.

Algorithm 2 presents the BTNN-based classification process, where a stacked ensemble integrates traditional ML and DL models to enhance recognition accuracy by capturing diverse patterns in environmental sound data.

Algorithm 2: Automatic environmental sound recognition using BTNN classifier

```
Input  : Number of features, training samples and testing samples
Output : Environmental sound classification
Initialize the population and fitness value
Define the set of matrix products of the C-path tensor M:
    $M(h_1,h_2,...,h_c) = J_1(:, h_1,:)J_2(:, h_2,:)...J_c(:, h_c,:)$
While do
    Compute magnitudes of each layer $A = \prod_{K=1}^{c} A_K G = \prod_{K=1}^{c} G_K$
    Compute TT-matrix factoring of matrix Z
        $Z(\mu_1(a),V_1(g),....,\mu_c(a),V_c(g)) = \prod_{K=1}^{c} J_K(:,\mu_K(a),V_K(g),:)$
```

```
6    Compute L-layer tensor neural network as $q \approx j\left(p \mid \{Z^{(L)}\}_{L=1}^{l}\right)$
7.   Update the fitness value
8.   Define Bayesian model $x\left(C \mid \{J_K\}_{K=1}^{c}\right) = \prod_{h=1}^{B} x(Q_h, j(p_h \mid [[J_1,...,J_c]]))$
7.   Compute absolute distribution for threshold set function
     $x(\theta \mid C) = \frac{x(C \mid \theta)x(\theta)}{x(C)} \propto x(C \mid \theta)x(\theta) = x(C,\theta)$
8.   End if
9.   Update the final value
10.  End
```

6 Results and discussion

This section presents and interprets the performance of the proposed environmental sound classification (ESC) system. We evaluate the model using three benchmark datasets—ESC-10, ESC-50, and UrbanSound8K—reporting standard metrics such as accuracy, precision, recall, and F1-score. Performance is analyzed using confusion matrices, per-class evaluation metrics, and ensemble configuration comparisons [38][39]. All experiments were conducted using Python 3.7, with TensorFlow 2.x and supporting scientific libraries. The Python version was selected due to compatibility requirements with pre-trained modules and BRSS optimization code.

6.1 ESC-10 performance

Figure 3 shows Confusion matrix for ESC-10 classification using the proposed SVM+BTNN ensemble model with fused spectrum features. As shown in Figure 2, the model achieves high per-class accuracy, particularly on classes such as 'rain' and 'clock_tick'. Minor confusions occur between acoustically similar classes like 'dog' and 'rooster'.

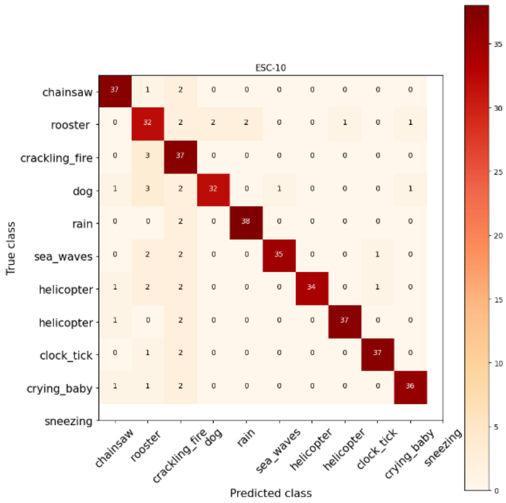

Figure 3: Confusion Matrix for ESC-10 dataset

6.2 ESC-50 performance

As shown in Figure 4, the proposed model performs reliably across diverse ESC-50 categories, with most classes achieving over 85% per-class accuracy. Misclassifications are concentrated among sound types with overlapping frequency textures, such as 'wind' and 'rain' or 'airplane' and 'thunderstorm', which suggests potential areas for future refinement.

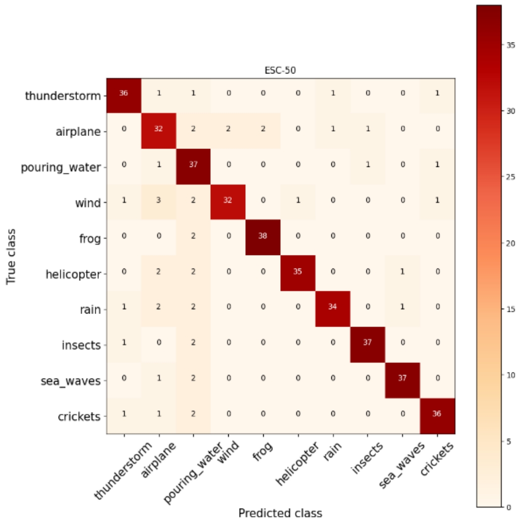

Figure 4: Confusion matrix for ESC-50 Dataset

6.3 UrbanSound8K performance

Figure 5. Confusion matrix of the UrbanSound8K classification task using the proposed SVM+BTNN model with fused features. The system achieves high accuracy on most urban sound classes, while exhibiting misclassifications primarily between structurally similar or overlapping sounds, such as “car horn” and “jackhammer.”

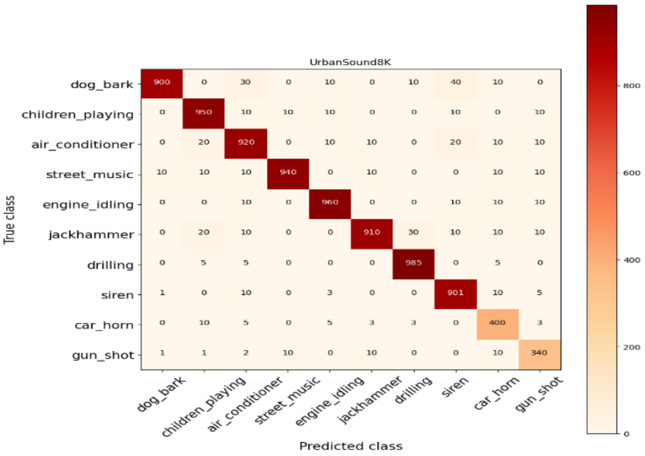

Figure 5: Confusion matrix for UrbanSound8K Dataset

6.4 Feature type comparison

Figure 6 shows that fused features obtained via BRSS significantly outperform individual spectral representations. The fused features achieve 98.98% accuracy, highlighting the benefit of combining complementary information from multiple spectrogram types.

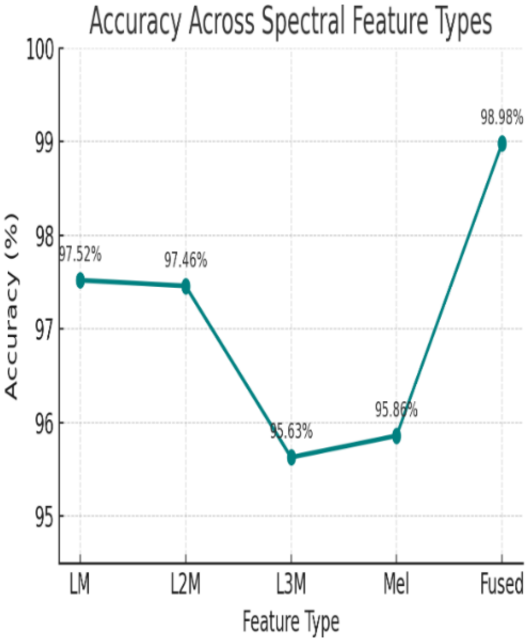

Figure 6: Accuracy across spectral feature types

6.5 Ensemble model comparison

Figure 8 compares the performance of three ensemble configurations. Among them, RF+BTNN achieves the highest accuracy (97.45%), demonstrating strong generalization and robustness

.

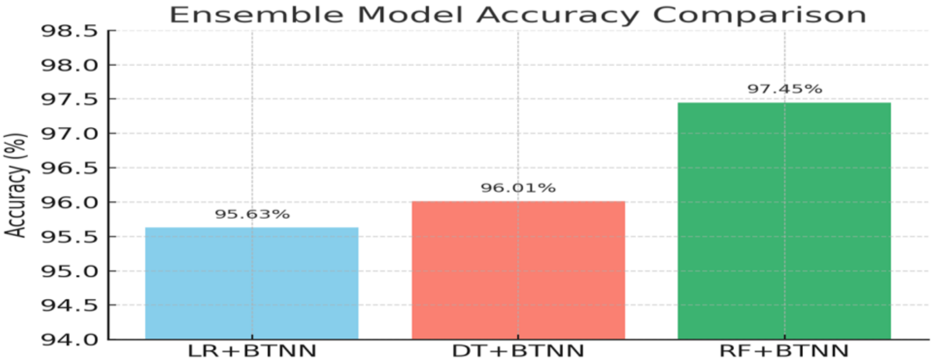

Figure 8: Ensemble model accuracy comparison

6.6 Comparative evaluation with existing models

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	46.05	45.76	45.11	45.43
Dilated CNN	52.02	51.73	51.08	51.40
EnvNet	57.99	57.70	57.04	57.37
CRNN	63.95	63.66	63.01	63.33
Dual ResNet	69.92	69.63	68.98	69.30
LR	75.24	74.95	74.30	74.63
DT	80.57	80.28	79.63	79.95
RF	85.89	85.61	84.95	85.28
SVM	91.22	90.93	90.28	90.60
LR + BTNN	96.55	96.26	95.61	95.93
DT + BTNN	96.87	96.58	95.93	96.26
RF + BTNN	97.20	96.91	96.26	96.58
SVM + BTNN	**97.53**	**97.24**	**96.58**	**96.91**

Table 2: Comparative analysis of existing and proposed methods using LM spectrum features

To further evaluate the superiority of our proposed ensemble framework, we conducted a detailed comparison using LM (Log-Mel) spectrum features, as shown in Table 2. The results highlight the performance of baseline deep learning models (CNN, CRNN, Dual ResNet), conventional machine learning classifiers (LR, DT, RF, SVM), and our proposed two-level stacked ensembles (ML + BTNN) [40][41]. Notably, the SVM+BTNN configuration outperforms all other models, achieving the highest accuracy of 97.53%, precision of 97.24%, recall of 96.58%, and F1-score of 96.91%. These results confirm that the integration of BTNN as a meta-classifier significantly boosts classification performance, especially when paired with robust feature fusion.

6.7. Per-class performance analysis

Figure 9 presents the per-class F1-scores for RF+BTNN. Most values exceed 0.98, with minimal variance between classes, indicating consistency

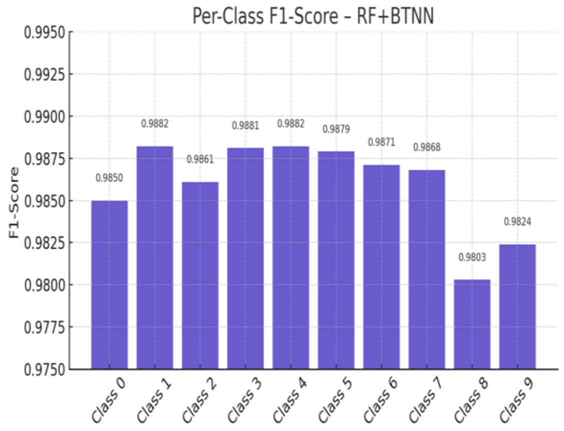

Figure 9: Per-Class F1-Score – RF+BTNN

Figure 8 provide additional insights into class-wise precision and recall, confirming that RF+BTNN maintains high performance across diverse acoustic environments.

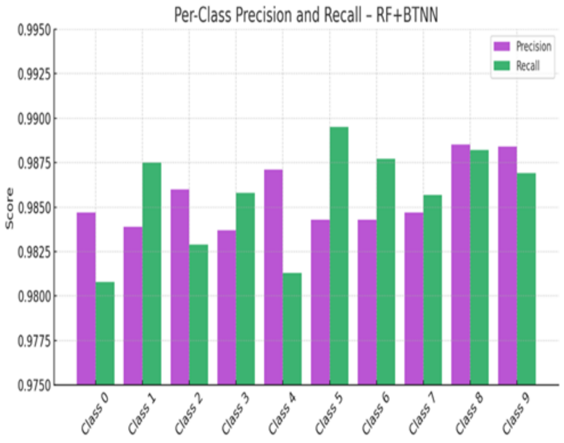

Figure 10: Per-Class Precision and Recall Bar chart for RF+BTNN

6.8. Summary Metrics

Table 3 summarizes the overall classification performance of the best-performing model (RF+BTNN) across the combined dataset.

Table 3: Overall evaluation metrics (combined datasets)

Metric	**Score**
Accuracy	93.30%
Precision	93.66%
Recall	94.55%
F1-Score	93.69%

6.9. Discussion

The proposed model demonstrates high accuracy and generalization across curated and real-world datasets. The BRSS-based fusion method enhances feature richness, while the RF+BTNN ensemble captures both linear and non-linear decision boundaries. The confusion matrices reveal most misclassifications occur in acoustically similar classes, suggesting room for improvement using temporal or attention-based modeling in future work. The per-class performance highlights the robustness of the system across minority and majority classes. The model maintains stable precision and recall even under class imbalance, validating its suitability for real-world deployment.

In terms of computational performance, the model processes samples with an average inference time of 8.3 ms, enabled by a compact BTNN with approximately 2.1 million parameters. This supports real-time ESC applications on resource-constrained devices. In conclusion, the proposed framework delivers robust and efficient environmental sound classification using stacked ensemble learning and optimal feature fusion. It outperforms baseline and traditional approaches, demonstrating clear applicability in domains like smart surveillance, wildlife monitoring, and ambient scene understanding.

7 Conclusion

This work presents a novel and efficient framework for automatic environmental sound classification that integrates BRSS-based optimal feature fusion with a stacked ensemble learning strategy. The system leverages DenseNet-161 to extract deep audio representations, which are then optimally combined using the Boosted Reptile Squirrel Search algorithm to reduce overfitting and improve generalization. Final classification is achieved through a two-level ensemble model comprising traditional classifiers (LR, DT, RF, SVM) and a Bayesian Tensorized Neural Network (BTNN) as a meta-learner. Experimental evaluations were conducted using ESC-10, ESC-50, and UrbanSound8K datasets. Our approach consistently outperformed baseline models across all datasets and feature types. Notably, the fused spectral features achieved the highest accuracy of 98.98%, confirming the advantage of multimodal fusion. Among ensemble configurations, RF+BTNN delivered the best results, demonstrating its ability to generalize across diverse sound classes. Compared to existing models such as Dual ResNet, the proposed system showed marked improvements in accuracy ranging from approximately 26% to 47% across different features. These gains emphasize the strength of combining lightweight deep feature extraction with optimization-driven fusion and a robust stacked ensemble. Overall, this work provides a scalable, accurate, and computationally efficient solution for environmental sound recognition and offers promising potential for deployment in edge devices, smart city infrastructure, and acoustic monitoring applications.

References

[1] M. Mirbeygi, A. Mahabadi, and A. Ranjbar. Speech and music separation approaches—A survey. *Multimedia Tools and Applications*, 81(15):21155–21197,2022. https://doi.org/10.1007/s11042-022-11994-1

[2] H. Kheddar, M. Hemis, and Y. Himeur. Automatic Speech Recognition using Advanced Deep Learning Approaches: A survey. *Information Fusion*, 109:102422, 2024. https://doi.org/10.1016/j.inffus.2024.102422

[3] R. Zaheer, I. Ahmad, D. Habibi, K. Y. Islam, and Q. V. Phung. A survey on artificial intelligence-based acoustic source identification. *IEEE Access*, 11:60078–60108,2023. https://doi.org/10.1109/ACCESS.2023.3283982

[4] Z. Chen, G. Xie, M. Chen, and H. Qiu. Model for Underwater Acoustic Target Recognition with Attention Mechanism Based on Residual Concatenate. *Journal of Marine Science and Engineering*,12(1):24,2024. https://doi.org/10.3390/jmse12010024

[5] H. Saini, N. Srinivasan, V. Šedajová, M. Majumder, D. P. Dubal, M. Otyepka, R. Zbořil, N. Kurra, R. A. Fischer, and K. Jayaramulu. Emerging MXene@ Metal–organic framework hybrids: design strategies toward versatile applications. *ACS Nano*, 15(12):18742–18776,2021. https://doi.org/10.1021/acsnano.1c06402

[6] V. Verma, A. Benjwal, A. Chhabra, S. K. Singh, S. Kumar, B. B. Gupta, V. Arya, and K. T. Chui. A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Scientific Reports*, 13(1):22719, 2023. https://doi.org/10.1038/s41598-023-49869-6

[7] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki. A Survey of Audio Classification Using Deep Learning. *IEEE Access*, 11:106620–106649, 2023. https://doi.org/10.1109/ACCESS.2023.3318015

[8] M. F. Siddique, Z. Ahmad, N. Ullah, and J. Kim. A Hybrid Deep Learning Approach: Integrating Short-Time Fourier Transform and Continuous Wavelet Transform for Improved Pipeline Leak Detection. *Sensors*, 23(19):8079, 2023. https://doi.org/10.3390/s23198079

[9] M. Tanveer, A. Rastogi, V. Paliwal, M. A. Ganaie, A. K. Malik, J. Del Ser, and C.-T. Lin. Ensemble deep learning in speech signal tasks: A review. *Neurocomputing*, 550:126436, 2023. https://doi.org/10.1016/j.neucom.2023.126436

[10] J. M. Navarro and A. Pita. Machine Learning Prediction of the Long-Term Environmental Acoustic Pattern of a City Location Using Short-Term Sound Pressure Level Measurements. *Applied Sciences*, 13(3):1613, 2023. https://doi.org/10.3390/app13031613

[11] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869, 2023. https://doi.org/10.1016/j.inffus.2023.101869

[12] A. Bansal and N. K. Garg. Robust technique for environmental sound classification using convolutional recurrent neural network. *Multimedia Tools and Applications*, 83(18):54755–54772, 2024. https://doi.org/10.1007/s11042-023-17066-2

[13] J. S. Coelho, M. R. Machado, M. Dutkiewicz, and R. O. Teloli. Data-driven machine learning for pattern recognition and detection of loosening torque in bolted joints. *Journal of the Brazilian Society of*

Mechanical Sciences and Engineering, 46(2):75, 2024. https://doi.org/10.1007/s40430-023-04628-6

[14] A. Mou and M. Milanova. Performance Analysis of Deep Learning Model-Compression Techniques for Audio Classification on Edge Devices. *Sci*, 6(2):21, 2024. https://doi.org/10.3390/sci6020021

[15] S. Sharma, K. Sato, and B. P. Gautam. A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques. *Sustainability*, 15(9):7128, 2023. https://doi.org/10.3390/su15097128

[16] S. B. Balasubramanian, P. Balaji, A. Munshi, W. Almukadi, T. N. Prabhu, K. Venkatachalam, and M. Abouhawwash. Machine learning based IoT system for secure traffic management and accident detection in smart cities. *PeerJ Computer Science*, 9:e1259, 2023.https://doi.org/10.7717/peerj-cs.1259

[17] S. Rani and G. Srivastava. Secure hierarchical fog computing-based architecture for Industry 5.0 using an attribute-based encryption scheme. *Expert Systems with Applications*, 235:121180, 2024. https://doi.org/10.1016/j.eswa.2023.121180

[18] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar, R. Ahmad, and A. H. Vakili. State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques. *Electric Power Systems Research*, 225:109792, 2023. https://doi.org/10.1016/j.epsr.2023.109792

[19] R. Moreno, F. Bianco, S. Carpita, A. Monticelli, L. Fredianelli, and G. Licitra. Adjusted controlled pass-by (CPB) method for urban road traffic noise assessment. *Sustainability*, 15(6):5340, 2023. https://doi.org/10.3390/su15065340

[20] K. Cao, T. Zhang, and J. Huang. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. *Scientific Reports*, 14(1):4890, 2024. https://doi.org/10.1038/s41598-024-55483-x

[21] M. Yildirim, S. Kiziloluk, S. Aslan, and E. Sert. A new hybrid approach based on AOA, CNN and feature fusion that can automatically diagnose Parkinson's disease from sound signals: PDD-AOA-CNN. *Signal, Image and Video Processing*, 18(2):1227–1240,2024. https://doi.org/10.1007/s11760-023-02826-2

[22] S. Mekruksavanich and A. Jitpattanakul. Hybrid convolution neural network with channel attention mechanism for sensor-based human activity recognition. *Scientific Reports*, 13(1):12067, 2023. https://doi.org/10.1038/s41598-023-39080-y

[23] S. Ansari, K. A. Alnajjar, T. Khater, S. Mahmoud, and A. Hussain. A robust hybrid neural network architecture for blind source separation of speech signals exploiting deep learning. *IEEE Access*, 11:100414–100437, 2023. https://doi.org/10.1109/ACCESS.2023.3318014

[24] X. Wang, Y. Wang, D. Liu, Y. Wang, and Z. Wang. Automated recognition of epilepsy from EEG signals using a combining space–time algorithm of CNN-LSTM. *Scientific Reports*, 13(1):14876, 2023. https://doi.org/10.1038/s41598-023-41537-z

[25] P. Rashmi and M. P. Singh. Convolution neural networks with hybrid feature extraction methods for classification of voice sound signals. *World Journal of Advanced Engineering Technology and Sciences*, 8(2):110–125, 2023. https://doi.org/10.30574/wjaets.2023.8.2.0083

[26] R. Jahangir. CNN-SCNet: A CNN net-based deep learning framework for infant cry detection in household setting. *Engineering Reports*, 6(6):e12786,2023. https://doi.org/10.1002/eng2.12786

[27] S. L. Ullo, S. K. Khare, V. Bajaj, and G. R. Sinha. Hybrid computerized method for environmental sound classification. *IEEE Access*, 8:124055–124065,2020. https://doi.org/10.1109/ACCESS.2020.3006082

[28] S. Liu, X. Fu, H. Xu, J. Zhang, A. Zhang, Q. Zhou, and H. Zhang. A fine-grained ship-radiated noise recognition system using deep hybrid neural networks with multi-scale features. *Remote Sensing*, 15(8):2068,2023. https://doi.org/10.3390/rs15082068

[29] X. Chen, Q. Yang, J. Wu, H. Li, and K. C. Tan. A hybrid neural coding approach for pattern recognition with spiking neural networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(5):3064–3078, 2024. https://doi.org/10.1109/TPAMI.2023.3339211

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

[31] D. Zhang, Z. Zhong, Y. Xia, Z. Wang, and W. Xiong. An automatic classification system for environmental sound in smart cities. *Sensors*, 23(15):6823,2023. https://doi.org/10.3390/s23156823

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[33] L. Abualigah, D. Yousri, M. Abd Elaziz, and M. A. A. Al-Qaness. Reptile search algorithm (RSA): A nature-inspired meta-heuristic optimizer. *Expert Systems with Applications*, 191:116158, 2022. https://doi.org/10.1016/j.eswa.2021.116158

[34] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov. Tensorizing neural networks. In *Advances in Neural Information Processing Systems*, 28:442–450, 2015. https://proceedings.neurips.cc/paper/2015/hash/6855456e2fe46a9d49d3d3af4f57443d-Paper.pdf

[35] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 37:1613–1622, 2015. https://proceedings.mlr.press/v37/blundell15.html

[36] M. Ye, X. Sheng, Y. Lu, G. Zhang, H. Chen, B. Jiang, S. Zou, and L. Dai. SA-FEM: Combined feature selection and feature fusion for students' performance prediction. *Sensors*, 22(22):8838, 2022. https://doi.org/10.3390/s22228838

[37] S. Z. Zhao, J. J. Liang, P. N. Suganthan, and M. F. Tasgetiren. Dynamic multi-swarm particle swarm optimizer with local search for large scale global optimization. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 3845–3852,2008. https://sci2s.ugr.es/sites/default/files/files/TematicWebSites/EAMHCO/contributionsCEC08/zhao08dms.pdf

[38] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *Proc. IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp.1–6,2015. https://doi.org/10.1109/MLSP.2015.7324337

[39] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian. Environmental sound classification with dilated convolutions. *Applied Acoustics*, 148:123–132,2019. https://doi.org/10.1016/j.apacoust.2018.12.019

[40] Y. Tokozume and T. Harada. Learning environmental sounds with end-to-end convolutional neural network. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2721–2725, 2017. https://doi.org/10.1109/ICASSP.2017.7952651

[41] H. Hojjati and N. Armanfard. Self-supervised acoustic anomaly detection via contrastive learning. In *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.3253–3257,2022. https://doi.org/10.1109/ICASSP43922.2022.9746207