iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76532

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

XGBoost and Pretrained PANN Embeddings on Strong Environmental Sound Event Classification

R K Santhia
Department of Computer Science, School of Engineering and Technology, Pondicherry University, India

Abstract: ***Sound Event Classification (SEC) is concerned with automatic recognition and classification of sounds in the environment which include alarms, animal sounds, human activities and machine sounds. SEC finds its applications in the real world in smart surveillance, urban sound, healthcare support system, and intelligent multimedia analysis. Nevertheless, traditional SEC methods using handcrafted features or convolutional neural networks (trained in their entirety) tend to be limited in the generalization ability as well as being highly sensitive to noise and overfit since labeled audio datasets are relatively small. To eliminate these constraints, the present work suggests a hybrid structure of sound event classification that gathers pretrained deep acoustic features extraction with ensemble learning. To produce deep acoustic embeddings with audio signals, a Pretrained Audio Neural Network (PANN) using CNN14 architecture trained on large-scale AudioSet is used as a fixed feature extractor to make robust deep acoustic representations of 2048 dimensions. These embeddings are then categorized into groups by an XGBoost classifier which is more adept at dealing with complicated decision boundaries and is more robust on small datasets. This workflow avoids the training deep networks entirely and overfitting is also drastically reduced. The suggested model is tested on ESC-50 environmental sound dataset, which is composed of 2000 audio samples divided into 50 sound classes and tested on the official 5-fold cross-validation setup. The results of the experiments show that the mean classification accuracy was 90.45, and the precision, recall, and F1-scores were similar in all the classes. The findings support the fact that the combination of pretrained acoustic representations with ensemble learning is an effective and credible solution to environmental sound event classification.***

Keywords: Pretrained Audio Neural Networks (PANN); Transfer learning; XGBoost Classifier; Deep Acoustic Embeddings; Ensemble Learning.

I. INTRODUCTION

Sound Event Classification (SEC) refers to the algorithmic recognition and arrangement of environmental sounds of audio files. Contrary to speech or music, sounds in the environment are very diverse, unstructured and in most cases they are unpredictable. These are a great variety of acoustic events that can be alarms, the sounds by animals, activities of humans, mechanical sounds, and the noise of nature. The main aim of SEC is to allow intelligent systems to sense and experience real world acoustics more or less like a human being does. Nevertheless, the high degrees of variability in the duration of sounds, the nature of frequencies, and intensity, the background noise, and the overlapping of the sources of sound are some of the complexities in sound event classification that makes research problematic. The reason why environmental sound recognition has gained relevance has to do with the vast number of applications that it has in the real world. SEC is applied to identify an abnormal or emergency event like a gunshot, glass breakage, and sirens in smart surveillance systems. SEC is used in urban sound monitoring applications to analyze noise to monitor traffic, building activity, and pollution of the environment. Sound event recognition in healthcare and assisted living settings can be used to aid the patient by monitoring falls, coughing, or sounds of distress. Also, SEC is important in indexing multimedia content, robotics and smart home systems, where audio-based context awareness can help to improve system intelligence and increase user engagement. Regardless of its significance and the general applicability of sound event classification, there are various limitations to the conventional approaches of sound event classification. The early SEC systems were based on handcrafted acoustic representations including Mel-Frequency Cepstral Coefficients (MFCCs), spectral representations and temporal representations with standard machine learned classifiers like Support Vector Machines, k-Nearest Neighbors, and Random Forests. These processes are strongly reliant on manual feature engineering and do not have strong resistance to noise and acoustic variations. Even though recent deep learning based methods have demonstrated better performance due to automatic learning of features based on time-frequency representations, they are often high computationally expensive and require large volumes of labeled data. Such models tend to experience overfitting and low generalization ability when used on smaller-size datasets, and more robust and data-efficient sound event classification strategies are required.

©IJRASET: All Rights are Reserved

II. RELATED WORK

Initial studies of sound event classification were based mostly on manual acoustic features with conventional machine learning algorithms. Such features as Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral roll-off, short-time energy, and zero-crossing rate were commonly used and were created to represent basic timefrequency properties of audio signals. Such characteristics were usually categorized with the help of algorithms like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Gaussian Mixture Models (GMM), and Random Forests. Although these methods were shown to have a reasonable level of performance on controlled datasets, feature engineering and parameter tuning were very important in determining their effectiveness. In addition, manually crafted attributes could not always reproduce intricate acoustic characteristics and were not as strong in noisy and realistic settings. As deep learning developed, one of the most popular methods of environmental sound classification is the convolutional neural networks (CNNs). CNN-based models are trained on discriminative features derived directly on timefrequency representations including spectrograms and log-Mel spectrograms, which does not require manual design of features. The multiple studies showed that deep CNN architectures are more effective as compared to traditional feature-based approaches on benchmark datasets like ESC-50 and UrbanSound8K. Nevertheless, training deep networks directly is expensive in both terms of size of labeled datasets and total computational resources. Deep learning models have been shown to be highly susceptible to overfitting and poor generalization when applied to relatively small datasets which restricts their practical use in data sparse situations. Transfer learning has been used to solve the issue of data scarcity in audio signal processing and to enhance generalization. Transfer learning methods utilize pre-trained models on large dataset, e.g. the AudioSet, and can apply them to tasks with small labeled datasets. It has been demonstrated that Pretrained Audio Neural Networks (PANNs), which are trained on a variety of environmental sounds, are highly able to extract high-level and transferable acoustic representations. Pretrained models can be utilized as fixed feature extractors to improve significantly in performance when used, and fine-tuning them to downstream tasks has allowed researchers to reduce training complexity. Transfer learning has therefore become one of the effective methods of improving the sound event classification performance in real world conditions.

Later hybrid methods which combine feature extraction by deep learning with classical or ensemble-based classifiers have been of interest. Under these techniques, the deep neural networks would be deployed to extract strong acoustic embeddings and therefore the classification would be done by machine learning models, and examples include Support Vector Machines, random forests, or gradient boosting algorithms. The ensemble classifiers especially those based on boosting have shown high performance with structured feature representation and enhanced resistance to overfitting. These hybrid models combine the representational capability of deep models with the decision-making capability of ensemble learning effectively, and such hybrid structures are appropriate in tasks of sound event classification that have limited data and complex acoustic patterns.

III. PROPOSED METHODOLOGY

This section outlines the environmental sound event proposal of the hybrid framework. The methodology combines trained deep extractions of acoustic features with ensemble based classification to attain strong results on small labeled data. The general steps in the workflow include audio preprocessing, extracting deep features with the help of a trained model, feature representation and final classification with the help of an ensemble learning algorithm.

A. Input Audio Data

The raw environmental sound records of the ESC-50 dataset are used as the starting point of the process. The audio files are brief.wav files that are the representation of one sound event of the animal sounds, human activities, or the sounds of the environment. These audio samples are the direct feeds to the proposed system.

B. Audio Preprocessing

The preprocess is done on each audio signal before the feature extraction to give uniformity. The audio files are converted to mono and resampled to 32 kHz sampling rate. The step is used to normalize the audio signals and guarantee their compatibility with the pretrained model. Preprocessing also eliminates variability due to various conditions of recording, as well as enhances consistency in feature extraction.

C. The next step is Pretrained Deep Feature Extraction (PANN – CNN14).

A Pretrained Audio Neural Network (PANN) which is an extension of CNN14 architecture is instead trained without training a deep neural network.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The model was trained with the large-scale AudioSet dataset that consists of millions of different sounds in the environment. Every preprocessed audio signal is sent to the CNN14 network and features are obtained at the embedding layer. Consequently, a deep acoustic embedding of dimensionality 2048 is obtained with respect to every sample of audio. Such embeddings have great temporal and spectral features of sound events.

D. Deep Acoustic Award Representation.

The resulting 2048 dimensional embeddings become a small and discriminative form of the original audio signals. These characteristics would describe key sound characteristics including frequency distribution, time characteristics and sound texture. Embeddings provide a powerful input to the classification by removing the manual feature design and offer a powerful input to machine learning.

E. Ensemble-Based Classification with XGBoost This stage is to classify with XGBoost using the ensembles.

The acoustic embeddings are profoundly deep fed into an XGBoost classifier to classify the final sound events. XGBoost is a gradient boosting ensemble algorithm which builds up a formidable family of decision trees into a robust classifier. It works well especially with organized feature data and medium size datasets. XGBoost detects complex decision boundaries on the deep embeddings thus boosting accuracy in classification and minimizing overfitting.

F. Strategy of Cross-Validation.

To guarantee trustful performance analysis, the model will be performed under the official 5-fold cross-validation protocol suggested by the ESC-50 dataset. The data is partitioned into five folds in this strategy. Training and testing on four folds and one fold respectively are performed in every iteration. Five repetitions of this process are done such that a single fold is utilized once as a test set.

G. Evaluation of performance.

The model performance is assessed based on the standard classification measures such as accuracy, precision, recall, and F1-score. These measures of evaluation are meant to give the overall picture of the effectiveness and strength of the model.

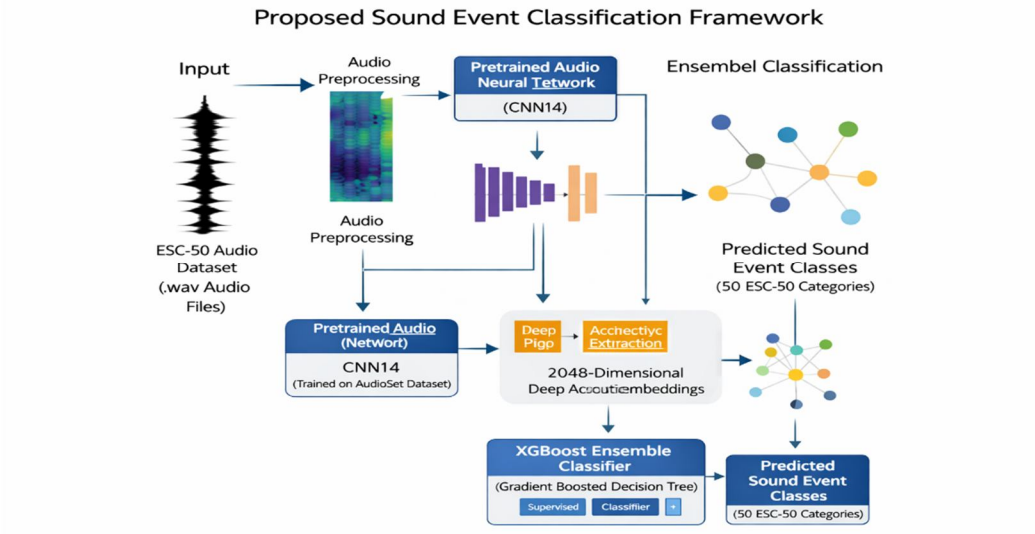


Figure 1: Architecture of Proposed Sound Event Classification Framework

IV. DATASET DESCRIPTION

The proposed ESC-50 dataset, which is a publicly available and well-known benchmark dataset on environmental sound event classification, is used to carry out the experiments in this study. The data set is very much tailored to aid in the testing of sound classification algorithms in factual acoustic settings. It includes brief recordings of environmental noise that represent a large selection of sound phenomena in the natural setting. ESC-50 dataset is well-curated and annotated and thus suitable.

TABLE I

Attribute	Description
Dataset Name	ESC-50 (Environmental Sound Classification)
Total Number of Audio Samples	2000
Number of Classes	50
Samples per Class	40
Audio Duration	5 seconds per sample
Audio Format	WAV (Uncompressed)
Sampling Rate	44.1 kHz (original), resampled to 32 kHz for experiments
Channels	Mono
Class Categories	Animal sounds, Human activities, Natural sounds, Interior and Exterior noises
Dataset Balance	Perfectly balanced
Evaluation Protocol	Official 5-fold cross-validation
Purpose	Environmental sound event classification

V. EVALUATION AND PERFORMANCE METRICS

A. Implementation Details

The proposed sound event classification scheme was done in Python programing language. The audio processing, feature extraction were performed through common audio analysis libraries, whereas the deep acoustic feature extraction was done through a pretrained Pretrained Audio Neural Network (PANN) with CNN14 architecture. The trained model was transformed into a generic feature extractor without further training and the ultimate classification step was done by the XGBoost ensemble learning algorithm. All the experiments have been done in a cloud-based setting and run on a CPU-based architecture. The training and testing process was conducted in strict compliance with the official 5-fold cross-validation procedure of ESC-50 dataset where four subsets were trained and the test was conducted with the rest of the subset. This was done five times such that a fold was employed as test set. Deep acoustic embeddings were also extracted on 2048 dimensions before the audio is classified on all audio samples. Each fold was trained and tested on the XGBoost classifier and the final performance was achieved by summing the results of each fold.

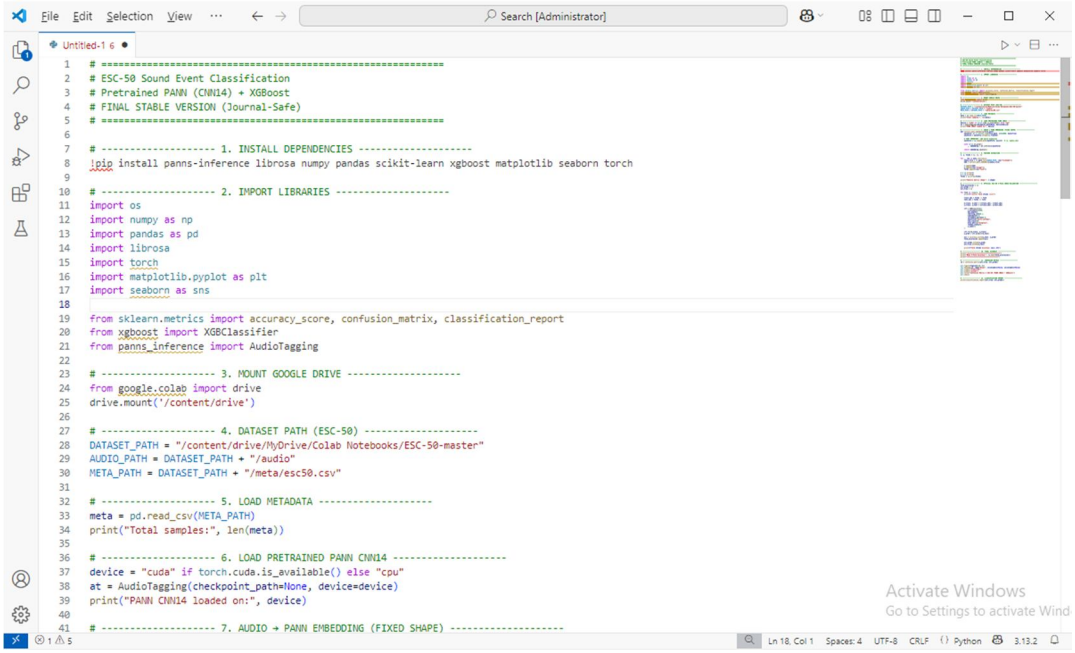


Figure 2: Implementation

B. Evaluation Metrics

The 5-fold cross-validation outcomes describe the way the suggested sound event classification framework works with various partitioning of the ESC-50 dataset and shape the foundation of the fold-wise accuracy graph. In Fold 1, the accuracy of the model is 0.9000, which means that the model has correctly classified 90 percent of the test samples, which is high level of baseline generalization. The best accuracy was obtained with Fold 2, with a result of 0.9250, thus demonstrating that this model actually worked quite well in this subset, and probably because there were clearer acoustic patterns or lesser inter-class overlap in the test data. Fold 3 accuracy reduced to a margin of 0.8975 which implies that there was marginally lower variation in the acoustic properties of the test samples still with competitive performance. Fold 4 achieved an accuracy of 0.9175, which validates that the model is always very generalized on various data splits. Fold 5 had the slowest accuracy of 0.8825, which suggested that this fold had difficulter samples or similar classes and hence caused more misclassifications. The 5-fold accuracy of 0.9045, which is the mean of the 5 folds, is the overall performance of the model. The fold-wise accuracy graph will illustrate these values, with every bar (or point) representing the accuracy of a certain fold, and it is possible to intuitively compare the performance consistency and show the strength of the proposed framework when working in a diverse range of test conditions.

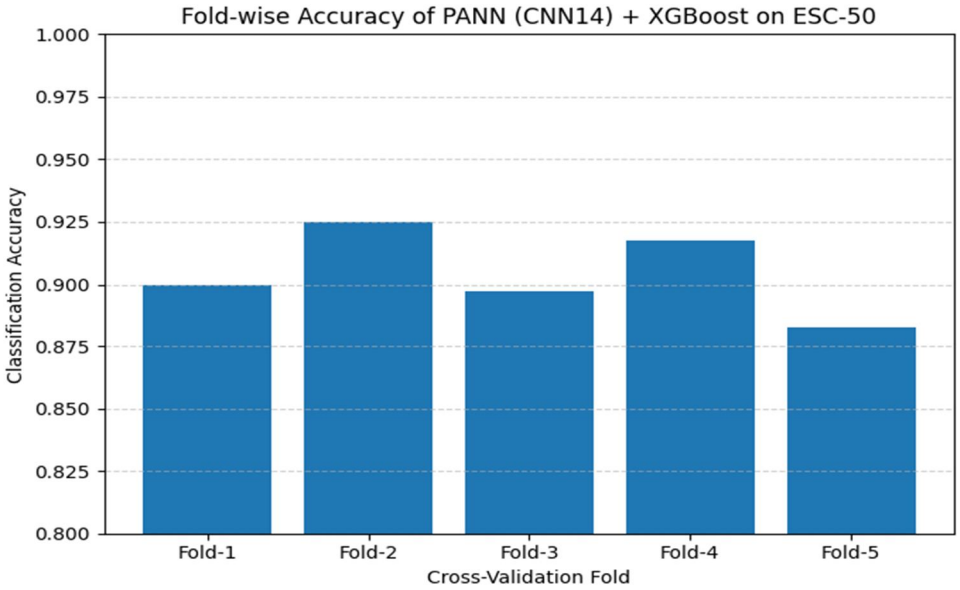


Figure 3: Fold-wise Accuracy of PANN(CNN14) + XGBoost

C. Overall Classification Accuracy

The main measure to be used in examining the performance of the proposed sound event classification framework was the overall classification accuracy. It is the proportion of audio samples correctly classified in the dataset to the number of samples in the dataset. This measure gives a worldwide estimate of the effectiveness of the model in all classes of sounds in the official 5-fold cross-validation regime.

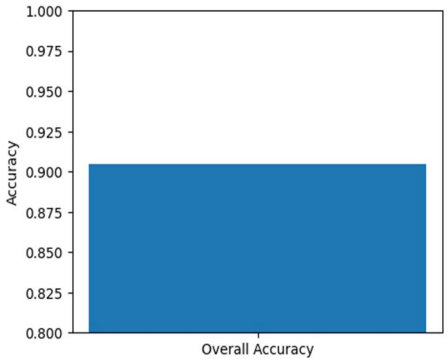


Figure 4: Overall Accuracy

D. Log-Mel Spectrogram

A spectrogram is an audio signal in time frequency form, indicating how the content of the frequency changes as time progresses. In log-Mel spectrogram, the intensity of color represents the sound energy and shows the acoustic patterns which can be used to classify sound events.

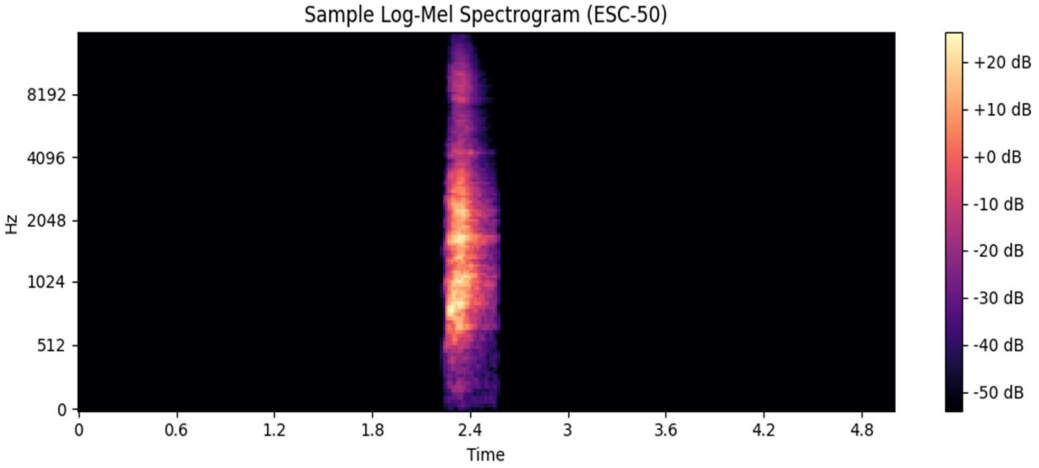


Figure 5: Log-Mel Spectrogram

E. Macro-Averaged Performance Metrics

The Macro-Averaged Metrics assess model performance by computing the metric on the individual classes and averaging the scores of all the classes. The method is equal treatment of classes irrespective of the number of samples in the class hence is particularly effective in evaluating balanced and fair performance of multi-class classification problems like environmental sound event classification.

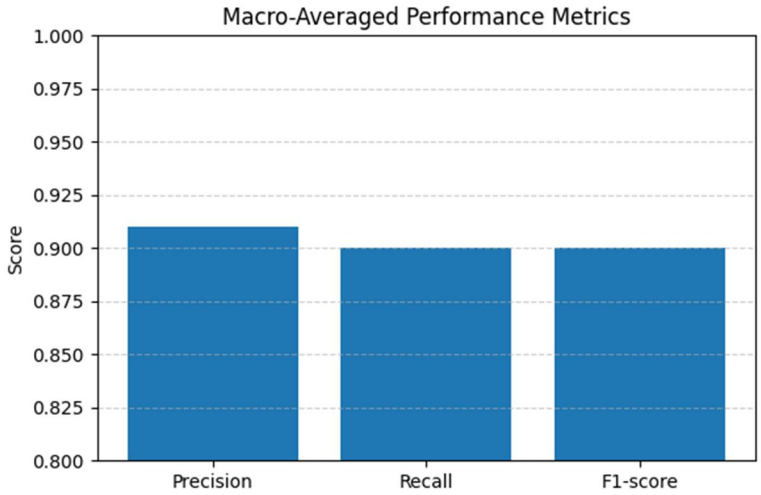


Figure 6: Macro-Averaged Performance Metrics

F. Precision

The macro-averaged precision is an average of the correctness of the model predictions in all classes of sounds. A large value of precision shows that on average, the model generates fewer false positive predictions of each class.

G. Recall

A macro-averaged recall is used to measure the capacity of the model to correctly recognize all the instances of every sound type. When the recall value is high it indicates that the model is able to capture truest sound events across classes.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

H. F1-score

F1-score is the harmonic mean of precision and recall, which is used to give the balance measure of the classification performance. The fact that its value is high demonstrates that the model has a good balance between false prediction rate and successful detection of sound events in all classes.

I. Weighted-Averaged Performance Metrics

The graph shows the weighted-average performance statistics of the proposed sound event classification framework. The value of 0.91 implies that there are correct predictions when considering the samples, and the values of recall and F1-score are close to 0.90, which is the indicator of successful detection and a reasonable trade-off between the precision and the recall, regarding the class frequency.

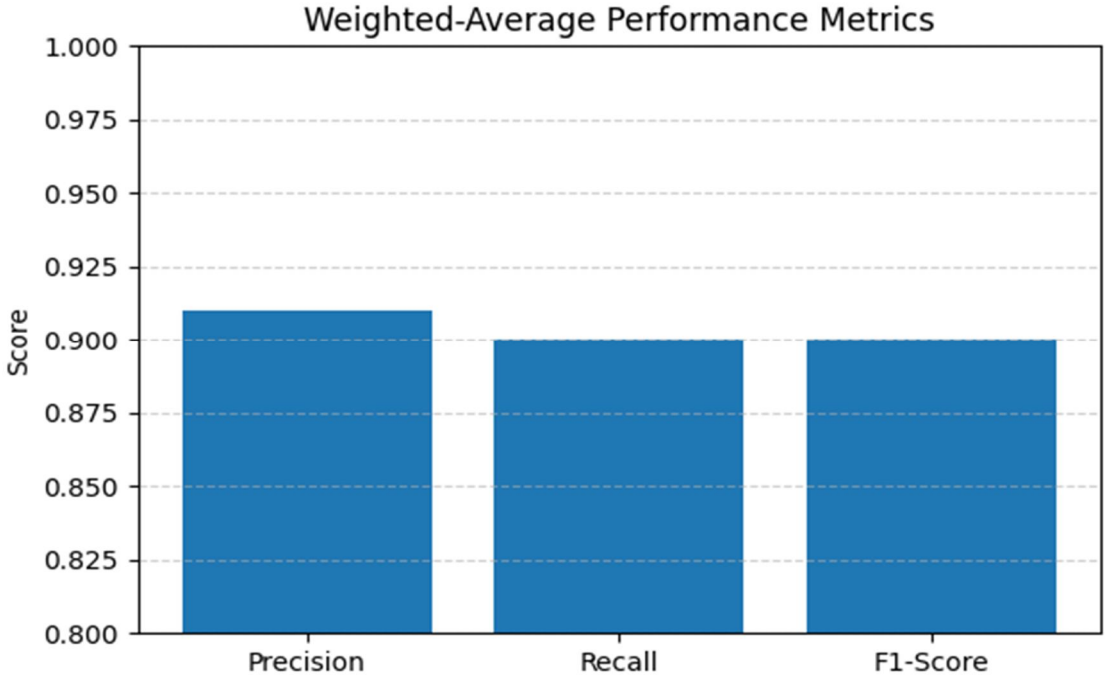


Figure 7: Macro-Averaged Performance Metrics

VI. CONCLUSION AND FUTURE WORK

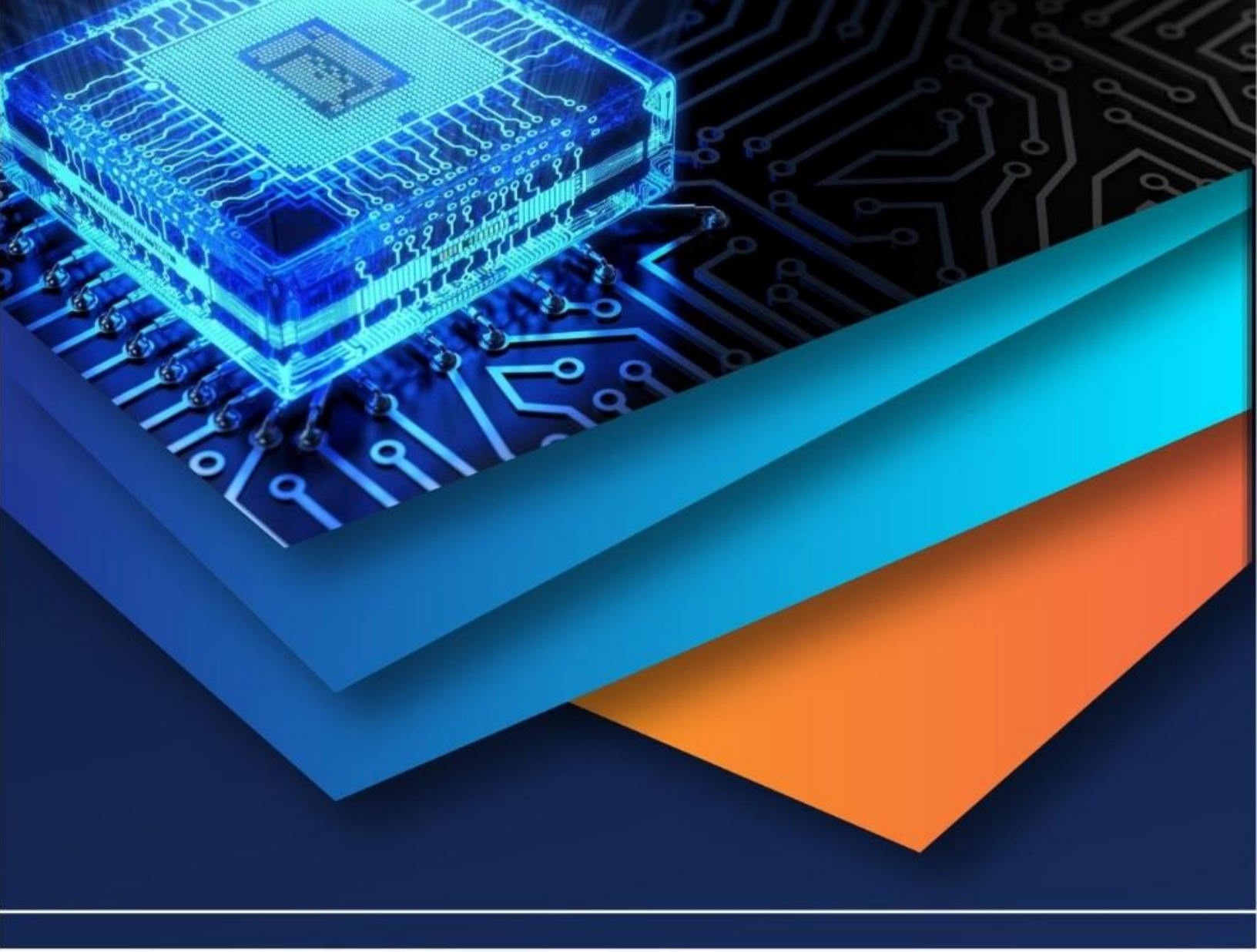
Environmental sound event classification is concerned with the automatic detection of sound events based on the signal of real-world audio. Environmental sounds are very heterogeneous, unorganized and most of the time are influenced by surrounding sounds and thus it is difficult to classify them correctly. Recent developments in deep learning have also enhanced performance through automatic learning of features using time-frequency representations, but there is high cost in terms of large labeled datasets and high computational cost to train deep models in mode. Transfer learning has therefore been introduced as one of the solutions to these challenges, by utilizing the benefit of using pretrained models on large-scale audio datasets. This paper is an experiment using Pretrained Audio Neural Network (PANN) as a CNN14 architecture to obtain the high-level acoustic embedding of an event, which reflects critical temporal and spectral features of sound events. Ensemble based XGBoost classifier is then used to classify these deep features which improves robustness and generalization.

REFERENCES

[1] V. Pann, K. S. Kwon, B. Kim, D. H. Jang, J. Kim, and J. B. Kim, “Robustness of CNN-based model assessment for pig vocalization classification across diverse acoustic environments,” Computers and Electronics in Agriculture, vol. 240, Art. no. 111181, 2026.
[2] A. Roy and U. Satija, “Effect of auscultation hindering noises on detection of adventitious respiratory sounds using pre-trained audio neural nets: A comprehensive study,” IEEE Transactions on Instrumentation and Measurement, 2025.
[3] S. Prakash and K. Sangeetha, “Systems classification of air pollutants using Adam optimized CNN with XGBoost feature selection,” Analog Integrated Circuits and Signal Processing, vol. 122, no. 3, p. 35, 2025.

[4] X. Zhou, B. Wang, X. Bao, H. Qi, Y. Peng, Z. Xu, and F. Zhang, "Quantitative detection of mixed gas infrared spectra based on joint SAE and PLS downscaling with XGBoost," Processes, vol. 13, no. 7, Art. no. 2112, 2025.
[5] S. Wan, S. Li, Z. Chen, and Y. Tang, "An ultrasonic-AI hybrid approach for predicting void defects in concrete-filled steel tubes via enhanced XGBoost with Bayesian optimization," Case Studies in Construction Materials, vol. 22, Art. no. e04359, 2025.
[6] A. Nogueira, H. Oliveira, J. Machado, and J. Tavares, "Sound classification and processing of urban environments: A systematic literature review," Sensors, vol. 22, 2022, doi: 10.3390/s22228608.
[7] A. Bansal and N. Garg, "Environmental sound classification: A descriptive review of the literature," Intelligent Systems with Applications, vol. 16, Art. no. 200115, 2022, doi: 10.1016/j.iswa.2022.200115.
[8] P. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," The Journal of the Acoustical Society of America, vol. 152, no. 1, p. 107, 2021, doi: 10.1121/10.0011809.
[9] P. Gairí, T. Pallejà, and M. Tresanchez, "Environmental sound recognition on embedded devices using deep learning: A review," Artificial Intelligence Review, vol. 58, 2025, doi: 10.1007/s10462-025-11106-z.
[10] M. Tailleur, P. Aumond, M. Lagrange, and V. Tourre, "Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network," The Journal of the Acoustical Society of America, vol. 156, no. 1, pp. 416–427, 2024, doi: 10.1121/10.0026479.
[11] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention-based convolutional neural network," Scientific Reports, vol. 11, 2021, doi: 10.1038/s41598-021-01045-4.
[12] A. Ekpezu, F. Katsriku, W. Yaokumah, and I. Wiafe, "The use of machine learning algorithms in the classification of sound: A systematic review," International Journal of Service Science, Management, Engineering, and Technology, vol. 13, pp. 1–28, 2022, doi: 10.4018/ijssmet.298667.
[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020, doi: 10.1109/TASLP.2020.3030497.
[14] O. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3978–3985, 2021, doi: 10.1109/TMM.2020.3035275.
[15] M. Tailleur, J. Lee, M. Lagrange, K. Choi, L. Heller, K. Imoto, and Y. Okamoto, "Correlation of Fréchet audio distance with human perception of environmental audio is embedding dependent," *arXiv preprint*, arXiv:2403.17508, 2024, doi: 10.48550/arXiv.2403.17508.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)