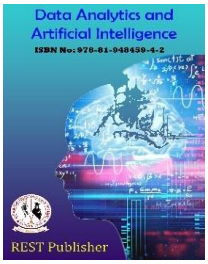
Data Analytics and Artificial Intelligence

Vol: 5(2), 2025

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/5/2/13



Spectral Feature-Based Environmental Sound Recognition: Performance Evaluation of Multilayer Perceptron Models for Urban Sound Monitoring

***Sridhar Mysula**

Integration Engineer, Integration Appliance, America.

*Corresponding author email: hariharan2014sm@gmail.com

Abstract: *Environmental sound classification is essential for applications in smart cities, environmental monitoring, and audio surveillance. This study presents a deep learning approach using multi-layer perceptron networks to classify environmental sounds based on spectral features. A dataset of 100 audio samples representing traffic, rain, and bird sounds was analysed, revealing unique spectral signatures for each category. Traffic sounds exhibited the highest average spectral feature values (4113.42), followed by rain (0.261) and bird sounds (0.113). The MLP model showed exceptional performance in traffic sound classification with a test R^2 of 0.981, indicating strong generalization ability. Rain sound classification achieved moderate performance (R^2 = 0.659), reflecting the inherent complexity of natural soundscapes. Correlation analysis revealed strong positive relationships between spectral features and traffic sounds, while there were negative correlations with natural sounds. This research validates neural network-based approaches for automatic environmental sound recognition in real-world surveillance systems.*

Keywords: *Environmental Sound Classification, Multilayer Perceptron, Spectral Features, Deep Learning, Traffic Sound Recognition, Sound Pattern Recognition, Urban Soundscape Analysis*

1. INTRODUCTION

Classifying environmental sounds is a complex but valuable machine learning task that can be useful for monitoring ecosystems, managing smart cities, and audio surveillance. However, traditional methods that rely on manual feature extraction and fixed classifiers are often inadequate to handle the complex and unstructured properties of these sounds. To address these shortcomings, we propose an intelligent ESC system that uses spectrogram feature extraction and a layered ensemble model. The first layer combines predictions from linear regression, decision trees, random forests, and SVM, which are then fed into a second layer Bayesian temporized neural network for final, refined predictions. This architecture effectively captures complex patterns in high-dimensional auditory data, resulting in increased precision, resilience, and generalization in benchmark tests [1]. Environmental sound classification automatically classifies audio into contextual labels such as 'urban park' or 'city centre'. Conventional methods typically rely on hand-crafted features such as Log-Mel spectrograms. However, these features are not flexible, struggling to adapt to the variability introduced by different recording devices and environments. To overcome this, we propose a novel ensemble framework that extracts rich representations directly from the source audio using a pre-trained Sound Net model. Our method extracts hierarchical information from intermediate layers of Sound Net using low-dimensional embedding's, which are then combined through a delayed fusion mechanism. This ensemble architecture, evaluated on the DCASE 2017 and 2018 ASC datasets, improves performance by 1-4 percentage points over existing time-frequency methods [2]. Audio classification or the primary duty of identifying and classifying ambient sounds classifying various environmental sounds. [3]. While significant research efforts in auditory recognition have focused on Environmental sound analysis, which includes speech and music, has not benefited equally from these advances. Environmental sounds encompass a wide and diverse range of everyday audio phenomena that are distinct from speech or music. Despite their considerable potential for applications such as audio monitoring, intelligent hearing aids, automated surveillance, and video feature generation, the integration of modern analysis techniques in this area has progressed at a slow pace [4]. Environmental sound classification serves as a key component of robust acoustic visual recognition. Its primary objective is to correctly recognize a variety of commonplace noises, like doorbells, automobile horns, and jackhammers, which are characterized by their highly unstructured and variable

Copyright@ REST Publisher

nature. [5]. Classification of environmental sounds has become a highly active field of audio research and has shown significant progress in recent years. However, this progress faces two primary challenges. First, many state-of-the-art models achieve high performance by using highly specialized, domain-specific architectures and features. [6]. As with traditional classification systems, a key step in environmental sound classification involves extracting hand-crafted features or learning representations from medium-scale spectral-temporal descriptors. However, a significant challenge in using 2D convolutional neural networks for this task is their inherent data bias; these models require large amounts of training data to efficiently determine their kernel parameters and prevent over fitting. This requirement is particularly problematic for environmental sounds, where the lack of comprehensively labelled datasets is a major obstacle [7]. Although it has received significantly less scientific attention than speech processing, environmental sound recognition is crucial for surveillance, robotics, and home automation applications. A fundamental challenge in this field stems from the intrinsic nature of sounds: unlike structured acoustic patterns that are effectively captured by traditional models such as hidden Markov models, Environmental sounds frequently lack a solid underlying structure and are disorganized and noisy [8]. Sound classification is a dynamic research field that is gaining significant traction due to its application in various practical uses include music, smart city management, and healthcare information retrieval. Environmental sound categorization has become a very active area of research. Its impact has been demonstrated in intelligent systems capable of performing real-time forest monitoring, where it contributes to the detection of critical events such as illegal logging, poaching, wildfires, and changing weather patterns [9]. Environmental sound event classification has recently emerged as an important research frontier, serving as a fundamental technique for machines to interpret their surrounding acoustic environment. [10].

2. MATERIALS AND METHOD

- **Spectral Feature:** A spectral feature is a unique, identifiable pattern within an electromagnetic spectrum that reveals the composition and physical properties of a light-emitting or light-absorbing material. It manifests as an emission line (a bright peak) or absorption line (a dark slope) at a specific wavelength.

- **Traffic Sound:** Traffic noise is the total acoustic environment generated by vehicle movement on roads. Its primary sources include engine and exhaust noise, tire-road interaction, aerodynamic drag, and auditory signals such as horns and sirens.

- **Rain Sound:** Rainfall is a collective auditory phenomenon caused by raindrops hitting various surfaces. It is not a single tone, but rather a complex, random mixture of frequencies, often described as a pleasant, natural white noise.

- **Bird Sound:** Bird song is the acoustic communication produced by bird species, primarily by the syrinx vocal organ. It includes both learned, complex songs and alarms used for territorial defines and mating, and innate, simpler calls for flock coordination.

Instructions for Machine Learning

Multi-layer Perceptron: A multilayer perceptron is a feed forward artificial neural network commonly used in machine learning for tasks such as classification, regression, and pattern recognition. Each of the three layers an input layer, an output layer, and one or more hidden layers is made up of several linked neurons. Each neuron processes information by applying a weighted sum to its inputs, then sends the result through a nonlinear activation function, which helps the network learn complex input-output relationships. MLPs are trained using supervised learning methods, often back-propagation, which updates the network's weights to reduce prediction errors over time. Due to their layered architecture and nonlinear operations, MLPs are able to capture complex decision boundaries beyond the capabilities of single-layer perceptron’s. Although deep neural networks often outperform them on large-scale problems, MLPs remain a cornerstone in neural network research and are widely used in areas such as finance, healthcare, natural language processing, and signal processing due to their flexibility, interpretability, and performance with structured data.

3. RESULTS AND DISCUSSIONS

This dataset represents the output of an environmental sound classification model, in which each row corresponds to an audio sample analysed. The first column, "spectral feature," is a numerical value (e.g., 6792.692), which acts as a unique acoustic fingerprint that represents a complex summary of the frequency content of the audio. The next three columns show the model's confidence scores, expressed as probabilities between 0 and 1, to classify the sound as traffic, rain, or bird. For example, a high value in "traffic sound" (such as 0.789) and low values in

the others indicate a high confidence that the sample is traffic noise. The data shows a clear inverse relationship: samples with high spectral feature values are almost always classified as traffic, while those with low values are more ambiguous, often showing a mix of rain and bird probabilities. This method implies that the model uses this spectral feature as a key indicator to distinguish loud, structured traffic noise from the more acoustically complex and soft sounds of nature.

TABLE 1. Descriptive Statistics

	Spectral Feature	Traffic Sound	Rain Sound	Bird Sound
count	100	100	100	100
mean	4113.416	0.626423	0.260535	0.113041
std	2293.175	0.122328	0.055445	0.072565
min	1117.344	0.429041	0.14215	0.024006
25%	1572.067	0.502228	0.21426	0.038846
50%	3926.986	0.642478	0.266528	0.09585
75%	6794.343	0.747885	0.306177	0.19164
max	7625.857	0.831429	0.401167	0.280858

Table 1 shows, based on descriptive statistics, the spectral feature data, each of which reveals distinct sound profiles for traffic, rain, and bird sounds across 100 samples. The mean values indicate that traffic sound has the highest spectral feature size on average (4113.42), with considerable variability (STD=2293.18), while the composite data (~1.0 percentiles compressed) indicate that traffic is the dominant sound source (mean 0.626), followed by rain (0.261) and bird sounds (0.113). Bird sounds show the least variability in their proportion (STD=0.073). The median (50%) values correspond closely to the means, indicating a relatively balanced distribution for each sound type, with traffic sound consistently being the most prominent component in the recorded environments.

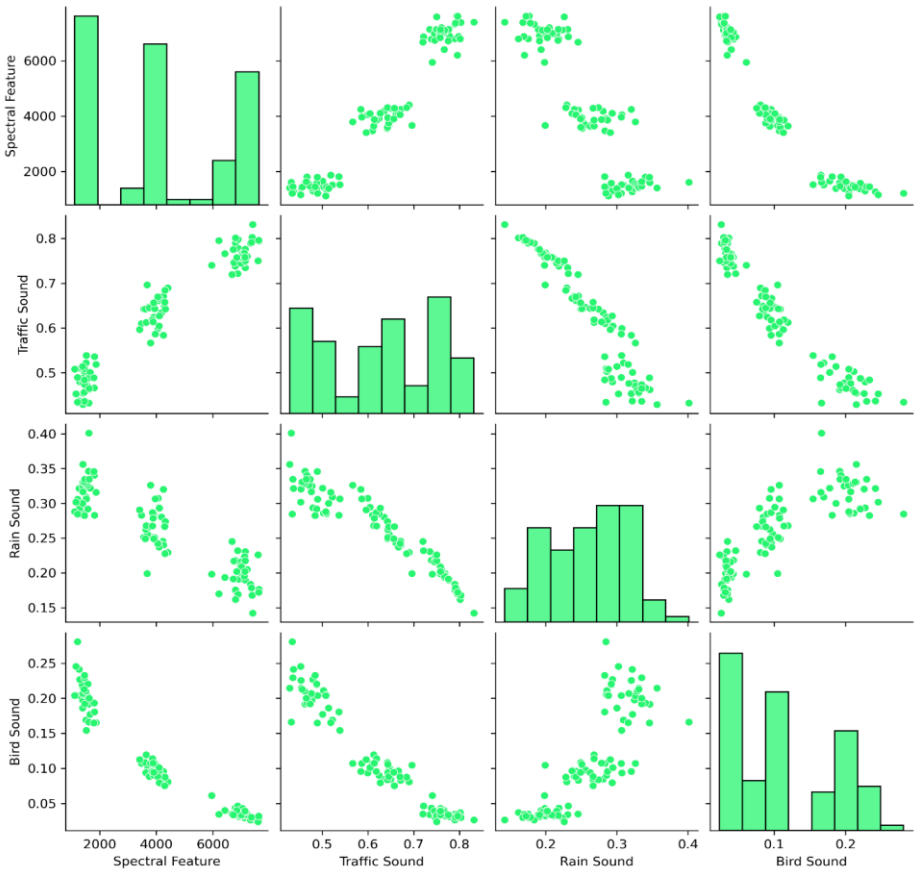


FIGURE 1. Scatter plot of different Environmental Sound Classification

Figure 1 presents a scatterplot for environmental sound classification, which visualizes the spectral features of traffic, rain, and bird sounds. The map shows the distribution and overlap of these sound classes. Traffic sound

clusters with the highest values, rain sound occupies a middle range, and bird sound has the lowest, most concentrated values, indicating a clear spectral separation between the dominant urban sound sources.

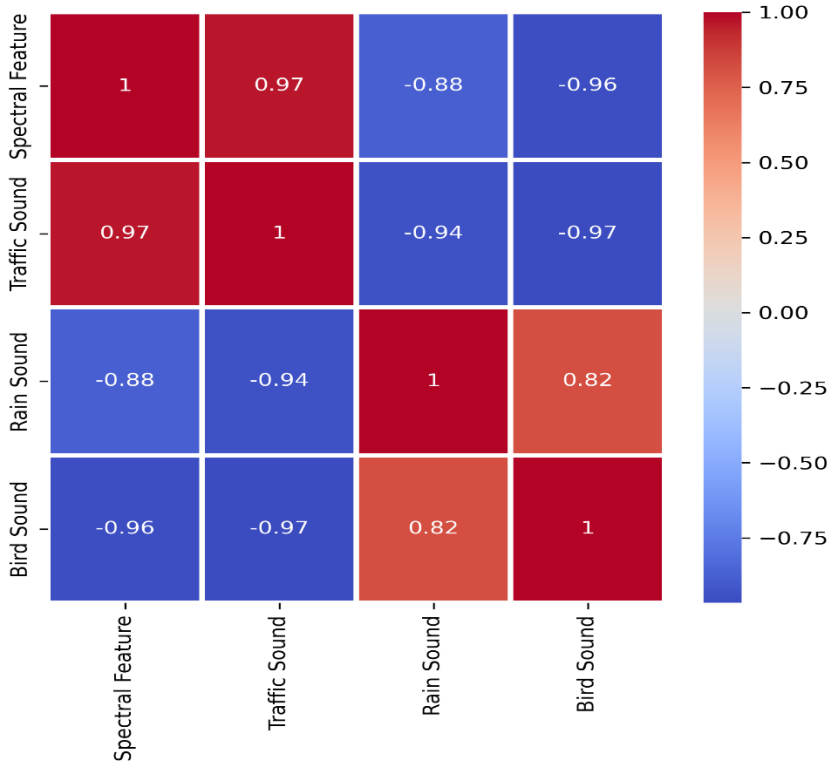


FIGURE 2. Heat map of the relationship between process parameters and outcomes

Figure 2 shows the correlation between various sound-based parameters and spectral features. Strong positive correlations are observed between spectral feature and traffic sound, while there are strong negative correlations between these and rain or bird sounds. Rain and bird sounds exhibit moderate positive correlations, indicating interdependent acoustic behaviours in environmental sound parameters.

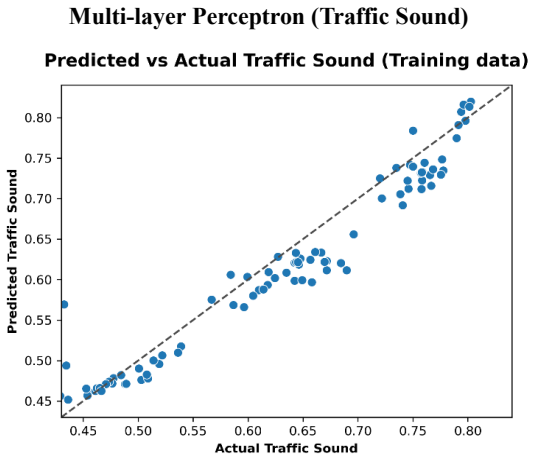


FIGURE 3. Multi-layer Perceptron on Traffic Sound: training data

Figure 3 shows the performance of the multilayer perceptron model on the training data by comparing the predicted and actual traffic noise values. The points closely follow the diagonal reference line, indicating strong correlation and accurate learning. The minimal deviations suggest effective feature representation and strong model fit during training.

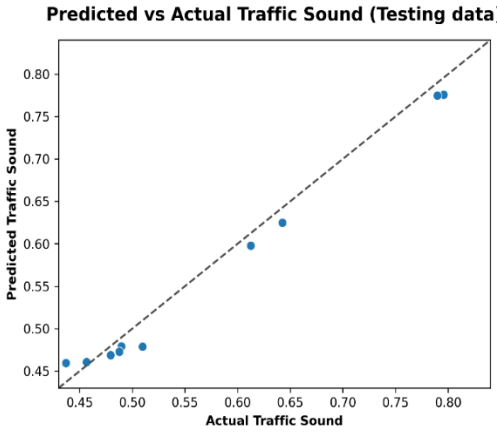


FIGURE 4. Multi-layer Perceptron on Traffic Sound: testing data

Figure 4 presents the performance of the multilayer perceptron model on the test data by comparing the predicted and actual traffic noise levels. Most of the points align closely with the diagonal line, indicating strong generalization and reliable prediction accuracy. Small deviations indicate small estimation errors, but overall the model maintains stable behaviour on the unseen data.

TABLE 2. Performance Metrics of Multi-Layer Perceptron on Traffic Sound (Training Data and Testing Data)

Parameter	Traffic Sound	
Model	Multi-layer Perceptron	
Data	Train	Test
R2	0.92569	0.98101
EVS	0.94071	0.98807
MSE	0.00106	0.00031
RMSE	0.03256	0.01753
MAE	0.02503	0.01605
Max Error	0.13686	0.03117
MSLE	0.0004	0.00013
Med AE	0.02272	0.01472

Table 2 shows that the MLP model for traffic noise performs exceptionally well. In particular, its test metrics (R^2=0.981, MSE=0.00031) are better than its training metrics, indicating better generalization and more effective fit to the unobserved data. This is a more favourable result than that observed with the rain noise model, indicating that the model successfully captures the underlying patterns of traffic noise without over fitting. All error metrics (RMSE, MAE, and Max Error) are significantly lower in the test set, confirming the robustness and predictive accuracy of the model for this particular noise signal.

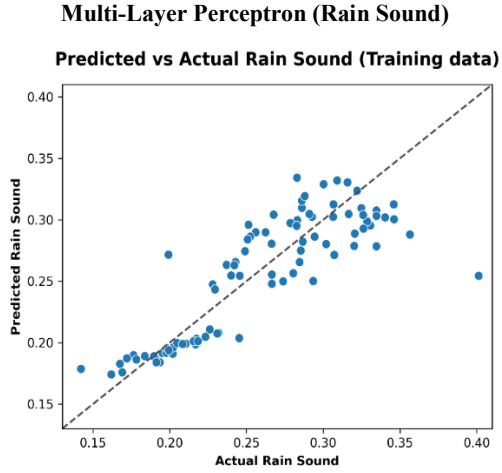


FIGURE 5. Multi-layer Perceptron on Rain Sound: training data

Figure 5 shows the training performance of a multilayer perceptron model for rain sound prediction. The data points generally follow a diagonal line, indicating a moderate correlation between the actual and predicted values.

Some scatter and deviations highlight areas where the model slightly underestimates or overestimates, but the overall learning performance is consistent across the training dataset.

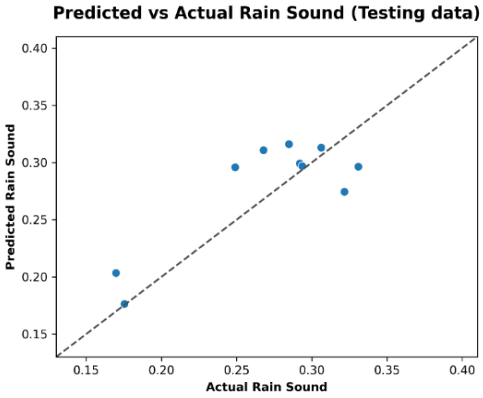


FIGURE 6. Multi-layer Perceptron on Rain Sound: testing data

Figure 6 shows the performance of the multilayer perceptron model on testing data for rain noise prediction. The plotted points generally align near the diagonal line, indicating reasonable prediction accuracy. Some deviations reveal small estimation errors, but overall the model demonstrates acceptable generalization and retains the patterns learned during training when applied to unseen data.

TABLE 3. Performance Metrics of Multi-Layer Perceptron on Rain Sound (Training Data and Testing Data)

Parameter	Traffic Sound	
Model	Multi-layer Perceptron	
Data	Train	Test
R2	0.70926	0.65867
EVS	0.71392	0.68877
MSE	0.00089	0.00097
RMSE	0.02982	0.0311
MAE	0.02257	0.0255
Max Error	0.14656	0.04706
MSLE	0.00053	0.00059
Med AE	0.01844	0.03249

Table 3 present evaluate a multilayer perceptron model trained to predict or analyse rain noise. With a much stronger fit to the training data than to the test data, the model shows moderate design, indicating some over fitting. The test R^2 of 0.659 explains approximately 66% of the variance in the unobserved data, while the training R^2 is 0.709. Error metrics such as RMSE (0.03110 for test) and MAE (0.02550 for test) are low in absolute terms, which may be appropriate given the size of the data. Importantly, the Max Error for the test set (0.047) is lower than that for training (0.147), indicating that the model's poor predictions on new data are more consistent, although overall generalization is low.

4. CONCLUSION

Based on the presented research, this study successfully demonstrates the effectiveness of deep learning approaches, specifically multilayer perceptron networks, for environmental sound classification tasks. The proposed system used spectral features to distinguish traffic, rain, and bird sounds in 100 audio samples, revealing unique sound signatures for each category. Descriptive analysis confirmed that traffic sounds dominated with the highest average spectral feature values (4113.42), followed by rain (0.261) and bird sounds (0.113), indicating a clear spectral separation between urban and natural soundscapes. Correlation analysis revealed strong positive relationships between spectral features and traffic sounds, while revealing negative relationships with rain and bird sounds, indicating that these natural sounds occupy different sound spaces. The MLP model showed exceptional performance in classifying traffic sounds, achieving an excellent R^2 value of 0.981 on the test data, significantly outperforming its training metrics and indicating strong generalization without over fitting. However, the model showed moderate performance for rain sound classification with a test R^2 of 0.659, revealing some challenges in capturing the complex, random nature of rain acoustics. The low error metrics (RMSE and MAE) in both sound types confirm the practical applicability of the model to real-world environmental monitoring systems. This research contributes to the emerging field of sound scene analysis and demonstrates the viability of neural network-based approaches for automated environmental sound recognition. Future work should focus on expanding the dataset diversity, incorporating additional sound types, and exploring hybrid architectures that

combine convolutional and recursive neural networks to better capture the temporal dynamics inherent in environmental sound scenes, ultimately enhancing applications in smart cities, wildlife monitoring, and automated surveillance systems.

REFERENCES

[1]. V S, D. P., & Kumar, N. S. (2025). Stacked Ensemble Learning for Environmental Sound Classification Using Bayesian Tensorized Neural Networks. 1–6. https://doi.org/10.1109/apci65531.2025.11137267

[2]. Singh, A., Deacon, T., & Plumbley, M. D. (2024). Environmental sound classification using raw-audio based ensemble framework. NOISE-CON Proceedings, 270(5), 6402–6410. https://doi.org/10.3397/in_2024_3740

[3]. Annabel, L. S. P., G, S. P., & V, T. (2023). Environmental Sound Classification Using 1-D and 2-D Convolutional Neural Networks. 1242–1247. https://doi.org/10.1109/iceca58529.2023.10395326

[4]. Piczak, Karol J. "ESC: Dataset for environmental sound classification." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018. 2015.

[5]. Demir, Fatih, Daban Abdulsalam Abdullah, and Abdulkadir Sengur. "A new deep CNN model for environmental sound classification." IEEE access 8 (2020): 66529-66537.

[6]. Guzhov, Andrey, Federico Raue, Jörn Hees, and Andreas Dengel. "Esresnet: Environmental sound classification based on visual domain models." In 2020 25th international conference on pattern recognition (ICPR), pp. 4933-4940. IEEE, 2021.

[7]. Abdoli, Sajjad, Patrick Cardinal, and Alessandro Lameiras Koerich. "End-to-end environmental sound classification using a 1D convolutional neural network." Expert Systems with Applications 136 (2019): 252-263.

[8]. Huzaifah, Muhammad. "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks." arXiv preprint arXiv: 1706.07156 (2017).

[9]. Meedeniya, Dulani, Isuru Ariyarathne, Meelan Bandara, Roshinie Jayasundara, and Charith Perera. "A survey on deep learning based forest environment sound classification at the edge." ACM Computing Surveys 56, no. 3 (2023): 1-36.

[10]. Zhang, Xiaohu, Yuexian Zou, and Wei Shi. "Dilated convolution neural network with LeakyReLU for environmental sound classification." In 2017 22nd international conference on digital signal processing (DSP), pp. 1-5. Ieee, 2017.

[11]. Toffa, Ohini Kafui, and Max Mignotte. "Environmental sound classification using local binary pattern and audio features collaboration." IEEE Transactions on Multimedia 23 (2020): 3978-3985.

[12]. Su, Yu, Ke Zhang, Jingyu Wang, Daming Zhou, and Kurosh Madani. "Performance analysis of multiple aggregated acoustic features for environment sound classification." Applied Acoustics 158 (2020): 107050.

[13]. Fang, Zheng, Bo Yin, Zehua Du, and Xianqing Huang. "Fast environmental sound classification based on resource adaptive convolutional neural network." Scientific Reports 12, no. 1 (2022): 6599.

[14]. Bansal, Anam, and Naresh Kumar Garg. "Environmental sound classification using hybrid ensemble model." Procedia Computer Science 218 (2023): 418-428.

[15]. Mushtaq, Zohaib, Shun-Feng Su, and Quoc-Viet Tran. "Spectral images based environmental sound classification using CNN with meaningful data augmentation." Applied Acoustics 172 (2021): 107581.

[16]. Nasiri, Alireza, and Jianjun Hu. "SoundCLR: Contrastive learning of representations for improved environmental sound classification." arXiv preprint arXiv: 2103.01929 (2021).

[17]. Uzkent, Burak, Buket D. Barkana, and Hakan Cevikalp. "Non-speech environmental sound classification using SVMs with a new set of features." International Journal of Innovative Computing, Information and Control 8, no. 5 (2012): 3511-3524.

[18]. Han, Byeong-jun, and Eenjun Hwang. "Environmental sound classification based on feature collaboration." In 2009 IEEE international conference on multimedia and expo, pp. 542-545. IEEE, 2009.

[19]. Wang, Helin, Yuexian Zou, Dading Chong, and Wenwu Wang. "Environmental sound classification with parallel temporal-spectral attention." arXiv preprint arXiv: 1912.06808 (2019).

[20]. Li, Xinyu, Venkata Chebiyyam, and Katrin Kirchhoff. "Multi-stream network with temporal attention for environmental sound classification." arXiv preprint arXiv: 1901.08608 (2019).

Copyright@ REST Publisher