Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33573

# *A CNN-Based Urban Sound Classification Method Using Mel Spectrograms for Noise Pollution Monitoring*

**Miaoke Li**

*Shanghai Pinghe School, Shanghai, China*
*maggielimiaoke@outlook.com*

***Abstract.*** Sound pollution is an easily overlooked source of pollution in cities, which can have an impact on people's physical and mental health, social relationships, and living environment. At the same time, alleviating urban noise pollution can also improve the ecological environment of cities. In order to alleviate urban noise pollution, this study designed a classification model that can accurately classify four different sounds: bird calls, children's playing sounds, air conditioning outdoor unit noise, and mechanical engine noise. The study used Mel spectrograms, which can convert the sound signals perceived by the human ear into spectrograms. Then, through the CNN convolution model, image features are processed, and different layers can identify different image features. This experiment also used the Edge Impulse platform, which optimized the development process of TinyML without the need to build a pipeline or transplant any models from scratch. The fully automated platform makes research more convenient and efficient. This model has achieved a relatively accurate sound classification task, with an accuracy of 88.4% and a loss of only 0.37.

***Keywords:*** Sound pollution, CNN, Mel spectrogram, edge impulse

## 1. Introduction

Sound pollution is always dismissed by humans. However, sound pollution has a huge impact on creatures' living conditions and physical and mental health. Specifically, for humans, severe sound pollution can cause mental illness. For instance, road rage. One of the main causes is the irritability caused by road noise. Thus, its impact is widespread, affecting humans' mental and physical health, community environment, ecosystem, habitats of animals, etc.

In the early field of noise pollution, the researchers were mostly focused on human life and health, a specific kind of sound, and methods to solve the problem before it happened. For instance, Cameron interviewed the citizens in a specific place to find out the association between noise and illness, the study also defined noise using its own methods [1]. Danilevičius' study can only classify specific, for instancecars [2]. But the type of sound it can distinguish was limited to Morillas' study [3]. Zambon's study focused on mapping, and volume monitoring [4]. Segura-Garcia's study mainly uses Raspberry Pi platforms and Tmote-Invent nodes to measure the equivalent noise pressure level [5].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33573

Different from the previous research, this study aims to optimize the living environment of all living organisms by training and identifying a large number of common sounds that are present in the daily life of various types, combined with existing volume recognition hardware devices, to assist urban management, reduce noise, and create an environment suitable for living that is conductive to physical and mental health. This study uses technical means to assist in and help with changing existing noise problems. Meanwhile, this study uses edge impulse as the training platform, which enables fully automated machine learning. Through the platform, the best parameters can be selected without the need for parameter tuning, resulting in excellent performance.

## 2. Method

### 2.1. Dataset preparation

The experimental data for this study come from some public datasets on the internet and individual sound data, specifically in the park. This study has 2 kinds of datasets: one is the public dataset, and the other is the personally collected dataset. The dataset has a total amount of 8019 data files, totaling 700MB, with a total duration of 3 hours, 44m, 0s, 3 hours, 44 minutes, and 0 seconds. For the public dataset, the study has selected the ESC-50 classic city dataset, zendodo's UrbanSound8K dataset, and the natural bird call dataset on the xeno -canto website as the public datasets to train the model. This study's own dataset comes from sound data collected at different times from four different parks around the communities, ensuring data diversity and experimental accuracy. This study used a classification method to process these data and divided them into five categories: air conditioning, bird chirping, human activities, mechanical sounds, and natural backgrounds, and labeled them accordingly for easier training. This paper uses Python to parse the file name and generate reusable code. By providing the CSV file with the specified file directory and output address, this paper can generate a TARGET_SR=16000 # sampling rate: 16000 Hz WINDOW_SIZE=16000 # window size: 1000ms (16000 Hz 1 s=16000 samples) WINDOW-STEP=8000 # sliding step size: 500ms (16000 Hz 0.5 s=8000 samples). The source data is cut by a sliding step size, and the threshold judgment of the maximum absolute amplitude (Peak Amplitude) is used to filter invalid slices. The original waveform data of the cut 1-second audio is analyzed, and the negative amplitude in the waveform is flipped to a positive number. The point with the loudest sound is found in this 1-second data. This not only processes the data structure, but also cleans the parts without sound. Prevent long speech from stuffing blank segments without bird calls into the model. Figure 1 is the interface of setting generate different sound features, showing the training set configuration parameters.

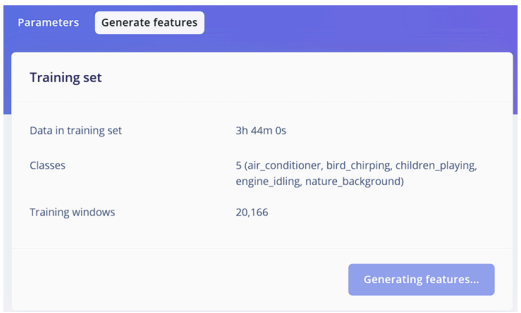


Figure 1. The interface of the setting generates different sound features, showing training set configuration parameters (picture credit: original)

DOI: 10.54254/2755-2721/2026.GU33573

## 2.2. Edge impulse-based classification

The study used Mel spectrogram and CNN convolution model to classify the data. Mel spectrogram is a tool that can convert sound into images, characterized by its ability to mimic the way the human ear listens to sound [6]. This article explains the basic logic of Mel Spectrogram and carefully explains how Mel Spectrogram works. The Mel spectrum is more sensitive to low-frequency sound perception and has the advantages of efficient and stable processing [7]. The CNN convolutional model is a deep learning algorithm that excels at processing data and images with grid structures [8-10]. It is specially designed for processing Grid-structured data. For instance, pictures and sound spectrum. Its core advantages are automatically extracting the layer's features and dramatically reducing the number of parameters. CNN structure has basically five layers: input layer, convolutional layer(Conv), pooling layer, fully connected layer(FC) and output layer (Softmax). The input later is used to normalize the input data and feed it into the networks. Conv is used to extract and learn the data through learnable filters. The activation layer is used to introduce the nonlinearity into the model. Pooling layer is used to decrease the computation cost. FC is used to flatten the high-dimensional features into a one-dimensional feature. The output layer converts the output into normalized probabilities for each category. Moreover, the CNN model is more convenient than other models, such as MFCC, because CNN can automatically learn the most discriminative data from the data without requiring too much manual adjustment.

Overall, the sound data is standardized using Java programs, and images are generated using Mel spectrograms. Then, a CNN convolutional model is constructed to analyze the sound.

This study uses the Edge Impulse platform in this experiment. Edge Impulse is a beginner-friendly platform that focuses on AI development. It provides a visual development environment, which is very convenient. At the same time, Edge Impulse provides a complete toolchain from data processing, features, model training, etc. Just input the data, and it can automatically learn and adjust, saving developers the trouble of building their own environment.

As shown in Figure 2 below, the study was trained for one hundred rounds with a learning rate of 0.002, and then set a 20% test set, with the rest being the platform's default data. But in order to improve the classification accuracy of this model, I manually annotated the data, filtered out nearly 50 erroneous audio, and ultimately achieved a good classification effect.

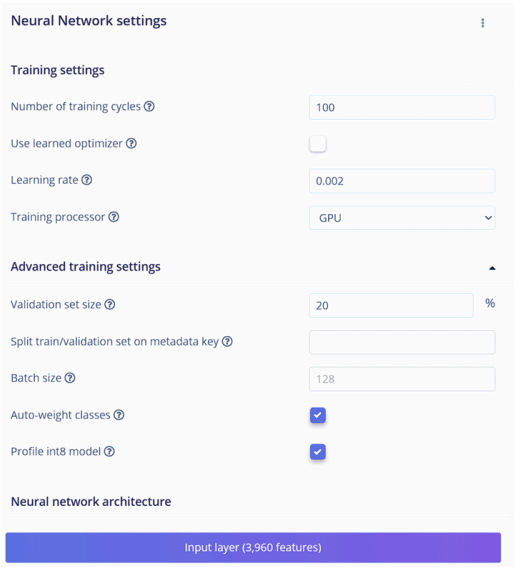


Figure 2. The neural networks settings (picture credit: original)

DOI: 10.54254/2755-2721/2026.GU33573

## 3. Results and discussions

Figure below shows the similarity of different sounds and the result of the experiment. The figure shows that the result of this study is valid. The samples are correctly classified as they formed some relatively independent clusters, which means features of different categories can be well distinguished.

It can be observed that on the graph that the misclassified data is mainly clustered in two partitions, forming two independent clusters. The children playing sound was misclassified into the sound features of the air conditioner, indicating that the features of these two types of sounds overlap and the model is difficult to distinguish.

According to the waveform diagram, it can be observed that there is no instantaneous pulse feature caused by children playing sound in the area selected by the gray box. The amplitude of the entire waveform is stable and the energy distribution is uniform, which is why the model cannot distinguish it. Figure 3. The result of the classifying process is shown by the distribution map for classifying four kinds of sound.




Figure 3. The result of the classifying process is shown by the distribution map for classifying four kinds of sound (picture credit: original)

This study used the Tiny Conv2D model with edge impulse in the experiment, and the final layer has 128 neurons and 0.5 dropout. The ability of this model to simulate human ear recognition is already very strong, and it can easily recognize the sounds emitted by different objects, which has achieved the purpose and effect required for the experiment very well.

According to Figure 4, the accuracy of the experiment was 88.4%, which indicates that the classification results of the experiment are generally good and can basically meet the goals and expectations of this experiment. The value of Loss is only 0.37, which is relatively low, indicating a good fit between the predicted results and the real situation, and sufficient convergence during training. According to the table in the figure, it can be observed that the accuracy of air conditioning noise and bird chirping is very high, at 95.4% and 95.8%, respectively, indicating that the model has

DOI: 10.54254/2755-2721/2026.GU33573

high feature recognition for such data. The accuracy of engine noise is second only to the first two, with only a small number of misjudgments. However, the sound of children playing is a clear weakness in this experiment, with an accuracy rate of only 74.0%. Among them, 15.8% of the samples were mistakenly identified as air conditioning noise. The results are consistent with the conclusion of the scatter plot above, and there is a high degree of overlap in the characteristics of the two sounds.

The value of Area under ROC Curve (AUC) is 0.98, which is very close to 1, indicating that the model has strong classification ability. The value of Weight average Precision is 0.9, indicating that the model's prediction credibility is high and the results are accurate. The weighted average recall value is 0.88, indicating that 88% of the models were correctly identified during testing, with a low rate of missed detections. The weighted average F1 score is 0.88, indicating that the overall classification performance of the model is excellent.

Overall, the experimental results were very accurate, with an accuracy rate of 88.4% and reduced data loss to 0.37. The running speed was almost twice as fast as other models. Although it requires a large amount of memory, it is still below the level of the same accuracy.

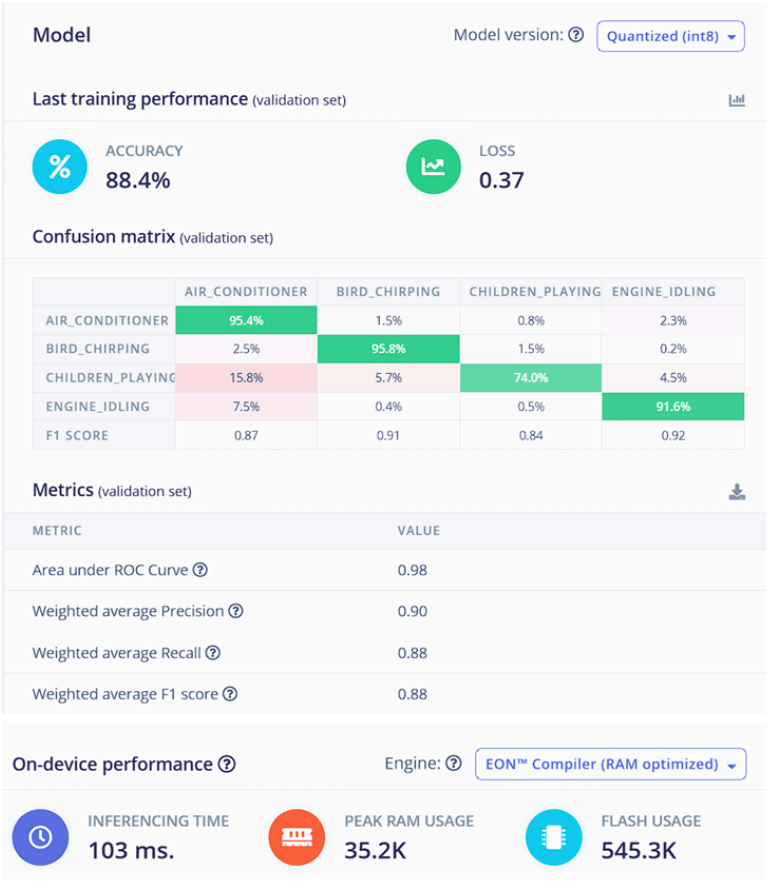


Figure 4. The result of experiment (picture credit: original)

## 4. Conclusion

This experiment mainly studied and designed a tool that can distinguish urban noise, which can reduce the impact of noise on people's living environment and physical and mental health, and protect the ecological environment of the city. The research mainly used CNN classification model

DOI: 10.54254/2755-2721/2026.GU33573

and Mel spectrogram. Through Python programming, the collected data was unified and processed. At the same time, the automated platform Edge Impulse was used to classify the data and establish a basic model. This model can classify bird calls, sounds of children playing, mechanical sounds, and air conditioning sounds. Although the sounds of children playing may be confused with bird calls, this model still achieves relatively accurate sound classification tasks. In the future, this model can be externally connected to existing volume recognition devices to display the size of sound while indicating the type of sound, which can help improve urban management and living environment.

## References

[1] Cameron, P., Robertson, D., & Zaks, J. (1972). Sound pollution, noise pollution, and health: Community parameters. Journal of Applied Psychology, 56(1), 67.
[2] Danilevičius, A., Karpenko, M., & Křivánek, V. (2023). Research on the noise pollution from different vehicle categories in the urban area. Transport, 38(1), 1-11.
[3] Morillas, J. M. B., Gozalo, G. R., González, D. M., Moraga, P. A., & Vílchez-Gómez, R. (2018). Noise pollution and urban planning. Current Pollution Reports, 4(3), 208-219.
[4] Zambon, G., Roman, H. E., Smiraglia, M., & Benocci, R. (2018). Monitoring and prediction of traffic noise in large urban areas. Applied Sciences, 8(2), 251.
[5] Segura-Garcia, J., Felici-Castell, S., Perez-Solano, J. J., Cobos, M., & Navarro, J. M. (2014). Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks. IEEE Sensors Journal, 15(2), 836-844.
[6] Imai, S., Sumita, K., & Furuichi, C. (1983). Mel log spectrum approximation (MLSA) filter for speech synthesis. Electronics and Communications in Japan (Part I: Communications), 66(2), 10-18.
[7] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), 53.
[8] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.
[9] Wu, J. (2017). Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 5(23), 495.
[10] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.