SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)
ISSN (Print) 1813-1743, ISSN (online) 2791-0547
Vol. 57: 2 (2025).
Doi: https://doi.org/10.26692/ surjss.v57i02.7708

Enhancing Urban Sound Classification with CNN-Transformer Hybrid Model and Spectrogram Augmentation

NOUMAN IJAZ[1], MD NAZMUl HASSAN[1], SANA ULLAH JAN[2], INSOO KOO[1]*

[1]Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, South Korea
[2]School of Computing, Engineering and Built Environment, Edinburgh Napier University, EH10 5DT Edinburgh, U.K.

***Cite this*:**
Nouman Ijaz et all. Enhancing Urban Sound Classification with CNN-Transformer Hybrid Model and Spectrogram Augmentation Sindh Uni. Res. J. (SS) 57. 02. 2025

Corresponding author
iskoo@ulsan.ac.kr

ABSTRACT

Urban Sound Classification (USC) is a crucial component of audio recognition systems, with applications in smart cities, surveillance, and multimedia. Despite significant advances, the classification of environmental sounds remains a challenge due to the complex nature of urban audio signals, characterized by high intra-class variability and overlapping sound events. In this paper, we propose a novel hybrid model that integrates the strengths of Convolutional Neural Networks (CNNs) and Transformer architectures to improve the identification accuracy of urban sounds. The CNN component effectively extracts local spectral features from Mel spectrograms, while the Transformer captures global temporal dependencies through self-attention mechanisms. Additionally, we incorporate advanced spectrogram augmentation techniques, such as time masking, frequency masking, and time warping, to further enhance the model's robustness and generalization capabilities. Experimental results on the UrbanSound8K dataset demonstrate that the proposed CNN- Transformer hybrid model outperforms traditional CNN and Long Short-Term Memory (LSTM)-based approaches, achieving a classification accuracy of 93.36%. These results highlight the effectiveness of combining CNNs with transformers and data augmentation strategies for robust urban sound classification.

Keywords: Urban sound classification, CNNs, LSTM, Spectrogram augmentation

1. INTRODUCTION

Environmental Sound Classification (ESC) is the task of identifying and classifying various types of environments based on the sounds, they produce [1]. The ability to accurately classify environmental sounds has become an increasingly important area of research within the field of audio recognition. This technology has been applied across a range of domains, including audio surveillance systems [2], [3], voice recognition for smart devices [4], [5], classification of mammals and bird sounds [6], [7], early detection of emergencies [8] and various gaming and multimedia applications [9], [10].

The process typically begins with the pre-processing of audio signals, followed by the extraction of relevant spectral features. In the final stage, these features are used for the classification of the audio signals.

Despite recent advancements in ESC, it remains a challenging problem that is far from fully resolved [11], [12]. We believe that the difficulty primarily stems from two key factors. First, environmental sounds exhibit highly complex and diverse characteristics [13]. The foreground and background sound events within an audio sample from the same class are often not identical, while samples from different classes frequently share similar sound events. This leads to significant intra-class variability and small inter-class differences. Second, the time-frequency characteristics of environmental sounds are inconsistent, with considerable variation in duration and spectral distribution. As a result, using a fixed scale for feature extraction and transformation may not be optimal for effective ESC.

Deep learning, specifically convolutional neural networks (CNNs), has shown promising solutions for urban sound classification [14]. CNNs excel at extracting local features from spectrograms, a two-dimensional representation of audio signals. However, their ability to capture long-range dependencies is limited [15]. Transformer architectures, originally developed for natural language processing, offer a potential solution by modeling global dependencies through self-attention mechanisms [16].

This research presents a hybrid CNN-Transformer model to overcome limitations of current techniques. The model combines the strengths of CNN for local feature extraction and Transformers for capturing global context. By combining these, we improved the accuracy and robustness of urban sound classification. To enhance the model's generalization, we employ spectrogram augmentation techniques, such as time warping and frequency/time masking. These augmentations replicate real- world fluctuations and distortions, enhancing the model's robustness to unseen data.

Experimental results on the UrbanSound8K dataset demonstrate the superior performance of our proposed model com- compared to traditional CNN-based methods. The hybrid CNN- Transformer architecture, combined with spectrogram augmentation, offers a promising approach for advancing urban sound classification and its applications in smart city technologies.

II. RELATED WORK

Urban sound classification has become a focus of research due to its potential applications in areas such as environmental monitoring, public safety, and smart city technologies. Early methods predominantly relied on traditional machine learning approaches, such as Support Vector Machines (SVMs) and Random Forests, which required handcrafted audio features for classification [17]. While these approaches provided initial insights, their performance was often constrained by their inability to generalize to the complex and diverse nature of urban soundscapes.

Deep learning methods, particularly Convolutional Neural Networks (CNNs), have significantly advanced the field by automating feature extraction. CNNs excel in processing spectrograms, a visual representation of audio signals, and extracting local spatial features crucial for sound classification. For instance, Salamon et al. [18] utilized CNNs in conjunction with Mel-spectrogram features to classify environmental sounds, demonstrating the efficacy of CNN-based approaches. Additionally, Recurrent Neural Networks (RNNs) [19], particularly Long Short-Term Memory (LSTM) networks [20], have been effective in capturing temporal dependencies in audio signals, addressing the sequential nature of sound events. However, both CNNs and LSTMs face challenges in modeling complex dependencies within the data, such as capturing relationships between long-range temporal patterns and distinguishing overlapping acoustic events [21].

To overcome these limitations, Transformer-based architectures have emerged as a promising solution. Originally developed for natural language processing tasks [22], Trans- formers employ self-attention mechanisms to capture global temporal dependencies, enabling them to model long-range relationships more effectively than traditional architectures [23]. Their application in audio classification has shown impressive results, with studies demonstrating their ability to complement CNNs by providing a more comprehensive representation of sound events. By combining the local feature extraction capabilities of CNNs with the global dependency modeling of Transformers, hybrid architectures have demonstrated superior performance in tasks requiring both spatial and temporal pattern recognition. The augmentation strategies used in our work align with recent trends in audio data enhancement. Techniques like time and frequency masking have been employed to improve model robustness in audio classification tasks [24].

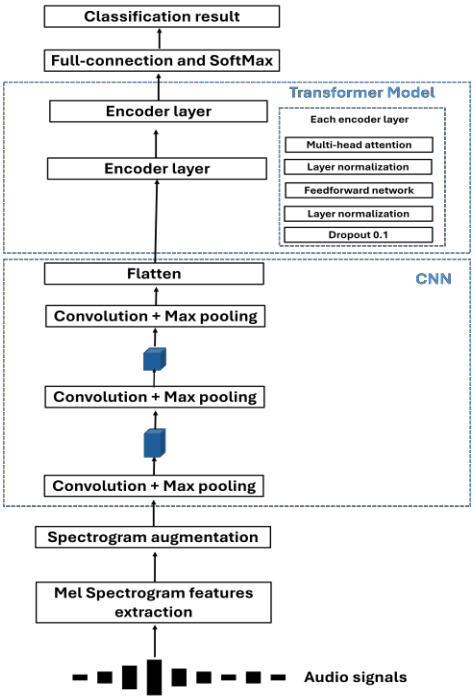

Fig. 1. Flowchart of the proposed method for urban sound classification.

III. PROPOSED METHOD

In this section, we describe the proposed method for urban sound classification, which involves a hybrid CNN- Transformer architecture combined with spectrogram augmentation techniques. The choice of a hybrid architecture stems from the goal of leveraging the distinct advantages of CNNs and Transformers. CNNs are highly effective in extracting localized spectral features from Mel spectrograms, making them well-suited for analyzing fine-grained patterns in audio data. On the other hand, transformers excel in modeling long-range temporal dependencies by utilizing self- attention mechanisms. By combining these two architectures, the hybrid approach enables comprehensive feature extraction, effectively addressing the individual limitations of CNNs and Transformers when used in isolation. The overall pipeline is illustrated in Figure 1, which outlines the flow of the model from input to output, as well as the data augmentation process.

A. DATASET

For our evaluation, we utilize the UrbanSound8K dataset [25], which comprises ten distinct urban environmental sound classes, such as air conditioners, dog barks, car horns, children playing, drilling, engine idling, gunshots, jackhammers, sirens, and street music. This dataset includes 8,732 labeled sound samples, each up to 4 seconds long, totaling approximately 9.7 hours of audio. Since the recordings were captured in real- world environments, many of the samples contain background

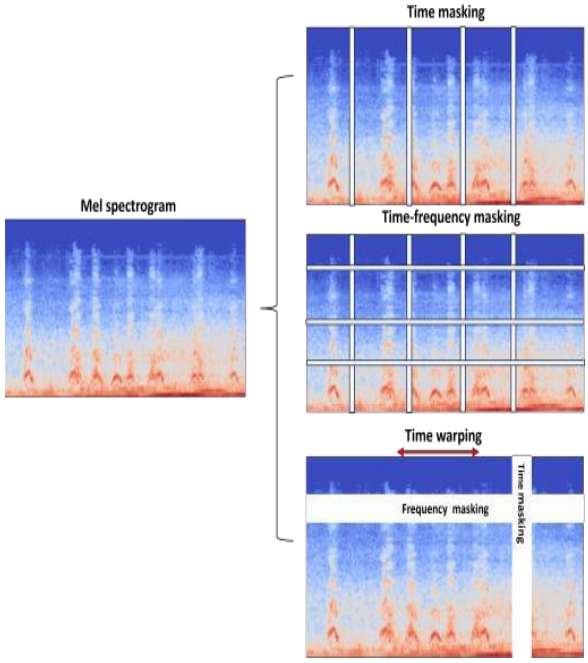

Fig. 2. Spectrogram augmentation techniques: time masking, frequency masking, and time warping.

noise or additional sounds besides the labeled one. The Ur- banSound8K dataset is organized into ten pre-arranged folds, ensuring that audio from the same recording is never present in both the training and testing sets.

B. PREPROCESSING AND FEATURES EXTRACTION

To analyze each audio file, we first transform it into a Mel spectrogram. This spectrogram is a visual representation that shows how the sound's energy is distributed over time and frequency. To create the Mel spectrogram, we break down the audio into small segments, calculate the power spectrum for each segment, and then map these values onto the Mel scale, which is a frequency scale that better reflects how humans perceive sound.

The Mel spectrograms we create have 64 frequency bands, and a varying number of time frames, depending on the length of the audio clip. We also apply logarithmic scaling to compress the dynamic range of

power values in the spectrogram. This makes it easier for the neural network to process the data.

C. SPECTROGRAM AUGMENTATION

To enhance the model's ability to handle diverse audio signals and prevent overfitting, we incorporate spectrogram augmentation techniques. These augmentations are applied directly to the Mel spectrograms and are based on the Spec Augment [26]. This method applies three operations: time masking, frequency masking, and time warping, as illustrated in Figure 2.

1) **Time Masking:** This technique involves selecting a random portion of the spectrogram, specifically a time segment between t_0 and $t_0 + t$ and setting its values to zero or a predefined constant. The length of this masked segment t is randomly chosen from a uniform distribution, with an upper limit defined by a maximum time mask parameter t_{max}. This method helps the model become more robust to small variations in timing, en- hancing its ability to handle sudden changes.

2) **Frequency Masking:** Similar to time masking, frequency masking operates along the frequency axis. A. random frequency range between f_0 and $f_0 + f$ is selected, and the values within this range are concealed. The width of the masked frequency band f is drawn from a uniform distribution with a maximum limit of f_{max}. This process helps reduce sensitivity to minor shifts in frequency, enabling the model to generalize better and ignore irrelevant frequency noise.

3) **Time Warping:** Time warping modifies the time axis of the spectrogram by either compressing or stretching it. An anchor point is selected at random along the time axis, and a warp factor is applied to both sides of this point, which shifts the spectrogram values accordingly. This simulates slight changes in the speed of the audio and helps the model better recognize distortions in time. Collectively, these augmentation techniques increase the diversity of the training data and improve the model's ability to learn stable patterns, thus enhancing its overall performance.

D. HYBRID CNN-TRANSFORMER MODEL ARCHITECTURE

The core of our proposed method is a hybrid model that combines Convolutional Neural Networks (CNNs) and Trans- former layers. The architecture is designed to leverage the strengths of both networks, CNNs for local feature extraction and Transformers for capturing global dependencies. The initial stage of our model involves feature extraction using convolutional neural networks (CNNs). This component is designed to identify local patterns within the Mel spectrograms. Three convolutional layers, each followed by batch normalization, max pooling, and dropout are employed to extract progressively higher-level features. The output of the CNN layers is then flattened into a one-dimensional vector. The flattened feature vector is subsequently fed into two Transformer encoder layers. These layers leverage self- attention mechanisms to capture global dependencies and contextual information within the input sequence. Each Trans- former layer consists of a multi-head attention module and a feed-forward neural network. The multi-head attention module utilizes 4 attention heads, each designed to capture different relationships across time frames, allowing the model to attend to multiple parts of the input sequence simultaneously. The feed-forward neural network consists of a dense layer with ReLU activation, which introduces non-linearity, followed by layer normalization to stabilize and accelerate the training process by normalizing the output at each layer. The output from the Transformer layers is fed into a fully connected layer with 256 neurons. After that, it passes through a SoftMax layer, which generates a probability distribution over the ten different urban sound classes. The model summary, presented in Table I, provides an overview of the architecture, detailing the input and output shapes at each layer along with the number of parameters involved.

E. EXPERIMENTAL SETUP

All programming tasks were conducted using Python version 3.6 within the Visual Studio Code Integrated Development Environment (IDE). Several libraries were employed to implement algorithms, including Pandas, NumPy, and Matplotlib for data processing. The Sound File library handled audio file input and output, while Librosa was used to extract features by generating Mel spectrograms. The neural network was built using PyTorch, with performance metrics, such as confusion matrices, calculated through Scikit-learn. Additionally, Tqdm was used to display progress bars during data processing, ensuring transparency. The dataset was divided into 6,286 training samples, 699 validation samples, and 1,747 test samples, providing a well-structured approach for model training, validation, and evaluation. To enhance the model's robustness,

spectrogram augmentation techniques were applied.

The model was trained for 100 epochs with a batch size of 32, using the Adam optimizer with a learning rate of $1 \times 10—4$, while tracking both training and validation loss. Model performance on the test set was evaluated using accuracy, precision, recall, and F1-score, offering insights into classification capabilities.

Table I Model Summary of the Proposed CNN-Transformer Architecture

Layer (type)	Output Shape	Param #
CNN Transformer	[32, 10]	–
Conv2d: 1-1	[32, 32, 64, 64]	320
MaxPool2d: 1-2	[32, 32, 32, 32]	–
Conv2d: 1-3	[32, 64, 32, 32]	18,496
MaxPool2d: 1-4	[32, 64, 16, 16]	–
Conv2d: 1-5	[32, 128, 16, 16]	73,856
MaxPool2d: 1-6	[32, 128, 8, 8]	–
Flatten: 1-7	[32, 8192]	–
Linear: 1-8	[32, 256]	2,097,408
Transformer Encoder: 1-	[32, 1, 256]	–
Module List: 2-1	–	–
Transformer Encoder Layer: 3-1	[32, 1, 256]	1,315,072
TransformerEncoderLayer: 3-2	[32, 1, 256]	1,315,072
Linear: 1-10	[32, 10]	2,570
Total params		**4,822,794**
Trainable params		**4,822,794**
Non-trainable params		**0**
Total mult-adds		1.39
Input size (MB)		0.52
Forward/backward pass size (MB)		60.23
Params size (MB)		17.19
Estimated Total Size (MB)		77.94

IV. RESULTS

The effectiveness of the proposed CNN-Transformer hybrid model was assessed through comprehensive experiments, and the results are discussed in this section. The training and validation loss, along with accuracy curves across 100 epochs, are depicted in Figure 3, showcasing the model's learning dynamics. As seen in the figures, the proposed model exhibited steady convergence with a notable reduction in both training and validation losses, and a consistent increase in accuracy over time, suggesting the robustness of the training process.

While the model demonstrated strong performance, its computational complexity is inherently higher due to the added transformer layers. Training the model requires approximately 78 MB of memory, as detailed in Table I. While this is manageable for systems equipped with GPUs, deploying it in resource-constrained environments, such as edge devices, may necessitate further optimizations.

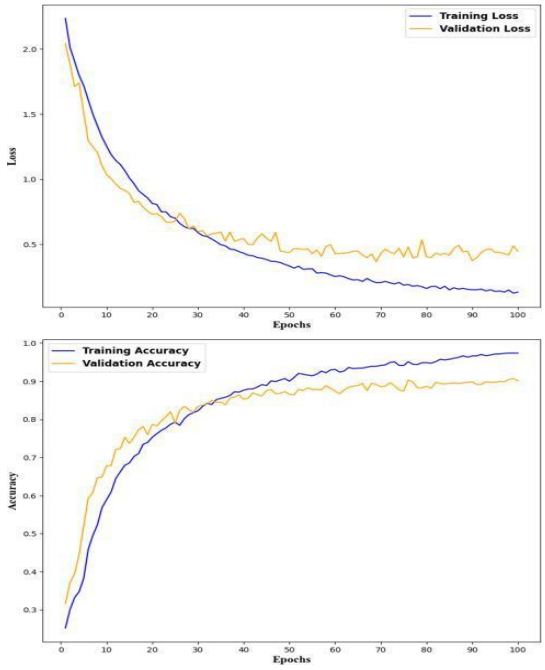

Fig. 3. Model Performance Evaluation

Table II Comparison of Methods Based on Performance Metrics

Method	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
CNN [27]	87.15	89.51	83.59	85.63
LSTM [27]	90.15	92.07	89.83	90.15
Transformer[28]	89.80	93.8	89.80	90.40
Proposed	93.36	94.80	93.81	93.70

To evaluate the performance of the proposed method, we conducted a comprehensive comparison with other established approaches, including CNN, LSTM, and a standalone Transformer model. The results of these models, along with the proposed method, are

summarized in Table II. As demonstrated, the proposed CNN-Transformer hybrid model achieved the highest accuracy of 93.36%, outperforming the CNN (87.15%), LSTM (90.15%), and Transformer (89.80%) models. Furthermore, the proposed method yielded superior precision, recall, and F-score metrics, emphasizing its capability to effectively handle the intricacies of urban sound classification.Notably, the integration of CNN and Transformer architectures allowed the model to leverage the strengths of both local feature extraction and global temporal dependency modeling,

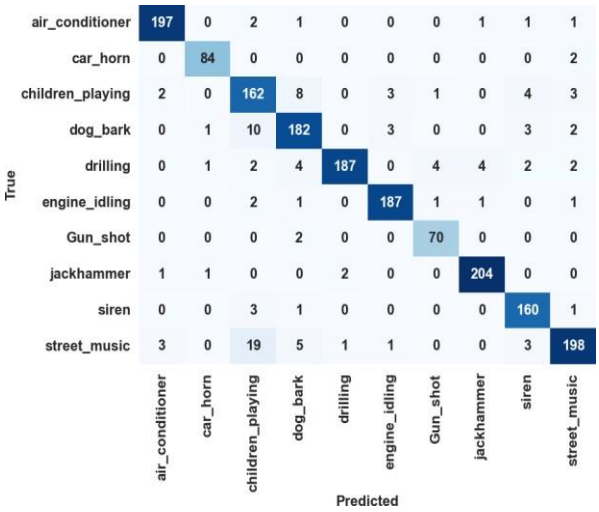

Fig. 4. Confusion Matrix from the Proposed Method.

Providing a significant advantage over standalone methods. Additionally, the use of data augmentation techniques, such as time masking and time warping, significantly contributed to the model's improved performance, increasing its accuracy from 92.16% to 93.36%. These results highlight the effectiveness of the proposed hybrid architecture and the augmentation strategies employed in this research.

The confusion matrix, as presented i n Figure 4 , provides a comprehensive evaluation of the CNN- transformer model's performance in classifying 10 urban sound categories. The model demonstrates significant success in distinguishing among various sound types, notably achieving high classification accuracy for classes such as "air conditioner"," engine idling," and "jackhammer". This is evidenced by the accurate identification of 197 "air conditioner" and 204 "jackhammer" instances. The minimal classification errors in these categories suggest that these sound classes exhibit distinct acoustic patterns, which the model is able to effectively capture and leverage during classification.

demonstrate that hybrid architecture significantly outperforms traditional methods, achieving a 93.36%

However, despite these successes, the confusion matrix also highlights areas where the model struggles, particularly in distinguishing between acoustically similar events. For example, "street music" was misclassified as "children playing" on 19 occasions. This misclassification likely stems from overlapping acoustic features, such as background noise and tonal variations, which can confuse the model in distinguishing between different urban sounds. Similarly, the misclassification of" dog bark" as" children playing" underscores the challenge of separating sound classes that share common auditory characteristics, such as pitch or intensity, which are difficult for even advanced models to distinguish with high precision. Despite these misclassifications, the overall distribution of the predictions, as illustrated by the diagonal elements of the confusion matrix, shows a strong alignment between true and predicted labels, reinforcing the effectiveness of the proposed model in this urban sound classification task. To evaluate the generalization capability of the proposed hybrid CNN-Transformer

Table III Performance of the Proposed Model on Urbansound8k Andesc-10 Datasets

Dataset	Accuracy (%)	Precision	Recall	F-score
ESC-10	88.75%	89.63%	88.75%	88.63%
Urbansound8K	93.36%	94.80%	93.81%	93.70%

model with spectrogram augmentation, we conducted experiments on the ESC-10 dataset [29] in addition to the UrbanSound8K dataset. ESC-10, an audio dataset with 10 environmental sound classes, provides a relevant benchmark to test the adaptability of the model to different audio environments. The results, summarized in Table III, demonstrate that the model achieves robust performance on both datasets, reinforcing its effectiveness in diverse scenarios.

V. CONCLUSION

In this paper, we presented a CNN-Transformer hybrid model designed to address the challenges of urban sound classification, leveraging both local feature extraction and global dependency modeling. By incorporating spectrogram augmentation techniques, the proposed approach enhances model robustness and generalization to real-world audio varia- tions. Our experimental results on the UrbanSound8K dataset

classification accuracy and offering improved precision, recall, and F1-score. While the model

successfully distinguishes between most urban sound classes, it still faces challenges in differentiating acoustically similar sounds, such as street music and children's play. Future research could explore further refinement in the model's attention mechanisms and introduce more advanced augmentation techniques to tackle these remaining challenges. Overall, the proposed hybrid model demonstrates strong potential for applications in smart city environments, public safety, and other urban sound analysis systems, while offering a promising foundation for future research and real-world deployments.

ACKNOWLEDGMENT

This result was supported by the "Regional Innovation System & Education (RISE)" through the Ulsan RISE Center funded by the Ministry of Education (MOE) and the Ulsan Metropolitan Government.

REFERENCES

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE.

Hassan, S. U., Khan, M. Z., Khan, M. U. G., & Saleem, S. (2019). Robust sound classification for surveillance using time frequency audio features. *In 2019 International Conference on Communication Technologies (ComTech)* (pp. 13–18). IEEE.

Li, Y., Li, X., Zhang, Y., Liu, M., & Wang, W. (2018). Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads. *IEEE Access, 6*, 58043–58055.

Tan, E.-L., Karnapi, F. A., Ng, L. J., Ooi, K., & Gan, W.-S. (2021). Extracting urban sound information for residential areas in smart cities using an end-to-end IoT system. *IEEE Internet of Things Journal, 8*(18),14308–4321.

Chandrakala, S., & Jayalakshmi, S. (2019). Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition. *IEEE Transactions on Multimedia, 22*(1), 3–14.

Ijaz, N., Banoori, F., & Koo, I. (2024). Reshaping bioacoustics event detection: Leveraging few-shot learning (FSL) with transductive inference and data augmentation. *Bioengineering, 11*(7).

Ijaz, N., Hasan, M. N., & Koo, I. (2025). Few-shot Transfer Learning Based Fault Classification in Wireless Sensor Networks. *IEEE Access*.

Shams, M. Y., Abd El-Hafeez, T., & Hassan, E. (2024). Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset. *Expert Systems with Applications, 249*, 123608.

Cohen-Hadria, A., Cartwright, M., McFee, B., & Bello, J. P. (2019).Voice anonymization in urban sound recordings. *In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia, 17*(10), 1733–1746.

Abeßer, J. (2020). A review of deep learning-based methods for acoustic scene classification. *Applied Sciences, 10*(6), 2020.

Nogueira, A. F. R., Oliveira, H. S., Machado, J. J., & Tavares, J. M. R. (2022). Sound classification and processing of urban environments: A systematic literature review. *Sensors, 22*(22), 8608.

Bansal, A., & Garg, N. K. (2022). Environmental sound classification: A descriptive review of the literature. *Intelligent Systems with Applications, 16*, 200115.

İnİk, Ö. (2023). CNN hyper-parameter optimization for environmental sound classification. *Applied Acoustics, 202*, 109168.

Knigge, D. M., Romero, D. W., Gu, A., Gavves, E., Bekkers, E. J., Tomczak, J. M., Hoogendoorn, M., & Sonke, J.-J. (2023). Modelling long range dependencies in N-D: From task-specific to a general purpose CNN. *arXiv preprint arXiv:2301.10540*.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 122666.

Agarwal, I., Yadav, P., Gupta, N., & Yadav, S. (2021). Urban sound classification using machine learning and neural networks. *In Proceedings of 6th International Conference on Recent Trends in Computing:* ICRTC 2020 (pp. 323–330). Springer.

Massoudi, M., Verma, S., & Jain, R. (2021). Urban sound classification using CNN. In 2021 6th International Conference on Inventive computation Technologies (ICICT) (pp. 583–589). IEEE.

Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. In 201826th European Signal Processing Conference (EUSIPCO) (pp. 2444–2448). IEEE.

Lezhenin, I., Bogach, N., & Pyshkin, E. (2019). Urban sound classification using long short-term memory neural network. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 57–60). IEEE.

Adavanne, S., Politis, A., Nikunen, J., & Virtanen, T. (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, 13(1), 34–48.

Rothman, D., & Gulli, A. (2021). Transformers for natural language processing (Vol. 267). Packt Publishing.

Zimerman, I., & Wolf, L. (2023). On the long-range abilities of transformers. arXiv preprint rXiv:2311.16620. Convolutional neural networks and data augmentation or environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283.

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia (pp. 1041–1044).

ACM. Park, D. S., Zhang, Y., Chiu, C.-C., Chen, Y., Li, B., Chan, W., Le, Q. V., & Wu, Y. (2020). Spec Augment on large scale datasets. In ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6879–6883). IEEE.

Bubashait, M., & Hewahi, N. (2021). Urban sound classification using DNN, CNN & LSTM: A comparative approach. In 2021 *International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies* (3ICT) (pp. 46–50). IEEE.

Nogueira, A. F. R., Oliveira, H. S., Machado, J. J., & Tavares, J. M. R. (2022). Transformers for urban sound classification---A comprehensive performance evaluation. Sensors, 22(22), 8874.

Piczak, K. J. (2015). ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM International Conference on Multimedia (pp. 1015–1018). AC.