A Comprehensive Approach to Urban Sound Detection with YAMNet and Bi-Directional LSTM

Tonghui Wu
Information and Computer Science, Xi'an Jiaotong-Liverpool University, Jiangsu, China

Keywords: Sound Event Detection and Classification, Deep Learning, Audio Signal Processing.

Abstract: Urban sound event detection and classification are increasingly critical in addressing the challenges posed by complex urban environments. As urbanization intensifies globally, traditional classification methods struggle with the overlapping sounds typical of cities. Leveraging advances in deep learning, this research aimed to enhance the accuracy of urban sound classification, which is essential for applications ranging from audio signal processing to noise monitoring to public safety. Utilizing the UrbanSound8K dataset, YAMNet, a pre-trained neural network, was combined with a custom Bidirectional LSTM network to develop a robust classification model. The model was evaluated using cross-validation, achieving a high Matthews Correlation Coefficient (MCC), indicating strong generalization to unseen data. Despite these positive outcomes, areas for further improvement were identified, particularly in distinguishing between acoustically similar sounds. This research contributes to advancing urban sound classification by integrating transfer learning and deep learning techniques, offering a solid foundation for future exploration in complex audio classification tasks and setting the stage for potential real-world applications.

1 INTRODUCTION

Urban sound detection and classification are critical tasks in audio processing, with applications ranging from noise monitoring to improving the quality of life in smart cities. As urbanization accelerates globally, cities are becoming denser and noisier, leading to increasingly complex sound environments (Nogueira et al., 2022). These environments are characterized by overlapping sounds from traffic, construction, and human activities, which pose challenges to traditional classification methods (Salamon & Bello, 2015). Effective sound classification is essential for developing responsive urban systems, such as real-time noise pollution monitoring and smart home devices that can recognize and respond to specific sounds. Accurate urban sound detection is also crucial for public safety, enabling the real-time identification of emergencies such as car accidents or gunshots. Therefore, advancing urban sound detection capabilities is both a technical challenge and a societal necessity, driving the demand for innovative approaches to manage the complexity and variability of urban soundscapes.

Sound event detection (SED) has been a key focus in audio processing, with several methods proposed to enhance both accuracy and efficiency. Heittola et al. emphasized the benefits of incorporating contextual information into automatic SED, significantly improving detection accuracy (Heittola et al., 2013). Cakir et al. presented multi-label deep neural networks for the identification of polyphonic sound events, signifying a notable increase in performance (Cakir et al., 2015). However, these methods still face challenges in polyphonic environments, where multiple sound events overlap. Recognizing this challenge, Mesaros et al. emphasized the need for robust evaluation metrics tailored to realistic scenarios, where multiple sound sources are active simultaneously (Mesaros et al., 2016). Parascandolo et al. explored RNN, particularly Bi-directional long short-term memory (Bi-LSTM) networks, to address polyphonic SED (Parascandolo et al., 2016). Expanding upon this work, Çakır et al. and Xu et al. integrated convolutional and recurrent neural networks (CRNNs) and proposed novel architectures such as gated convolutional neural networks (CNN) with temporal attention, which improved classification accuracy and earned recognition in challenges like DCASE 2017 (Çakır et al., 2017; Xu et al., 2017). Further extending these

Wu and T.
A Comprehensive Approach to Urban Sound Detection with YAMNet and Bi-Directional LSTM.
DOI: 10.5220/0013486700004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 23-29
ISBN: 978-989-758-754-2

ideas, Adavanne et al. developed CRNNs for joint acoustic event localization and identification in three-dimensional environments (Adavanne et al., 2018). Finally, Turpault et al. explored SED in domestic environments using weakly labeled data, significantly contributing to the field by introducing the DESED dataset (Turpault et al., 2019). While these studies have advanced the field considerably, they primarily focus on environments where sound sources are well-defined or isolated. Accurately detecting and classifying overlapping or noisy sound sources remains a significant challenge in current research. Furthermore, most existing models are heavily data-dependent and struggle to generalize to new, unseen urban soundscapes, particularly in highly variable and unpredictable urban environments. This gap underscores the need for models that not only perform well on standardized datasets like UrbanSound8K but also demonstrate robust generalization capabilities in real-world applications.

To address these limitations, this study investigates the application of transfer learning by utilizing YAMNet a pre-trained model designed for general audio classification tasks. The motivation behind this approach is to utilize YAMNet's robust feature extraction capabilities, which have been trained on a diverse range of sound categories, to optimize the precision and generalization of urban sound categorization models. This study also integrates YAMNet with a custom Bidirectional LSTM network to classify sounds from the UrbanSound8K dataset. Cross-validation approaches were employed to train and assess the model, hence assuring its robustness. Key preprocessing steps included waveform standardization and label encoding. Metrics such as accuracy, F1-score, and Matthews Correlation Coefficient (MCC) were used to evaluate the effectiveness of the model. By integrating YAMNet with a custom deep learning model, improved classification accuracy and generalization were achieved. This approach not only addresses existing gaps in urban sound classification but also opens new avenues for applying transfer learning to other complex audio classification tasks.

2 METHODOLOGY

This section details the steps taken in preparing the data, building the model, and optimizing it for urban sound classification. It covers the data preparation, feature extraction using YAMNet, the architecture of the Bi-Directional LSTM model, and the choice of loss function and optimizer. Figure 1 illustrates the sequential process of this study.

2.1 Data Preparation

This study utilizes the UrbanSound8K dataset, a widely recognized benchmark for urban sound classification tasks. The dataset comprises 8,732 labeled sound excerpts, each lasting up to 4 seconds. These excerpts are categorized into 10 classes commonly found in urban environments such as the street music, children playing, dog bark. These classes were chosen for their prevalence in urban noise complaints, making them highly relevant for urban sound detection studies. The distribution of the total occurrence duration per class and salience is illustrated in Figure 2(a). To ensure balanced class distribution, each class was limited to 1,000 slices, totaling 8,732 labeled slices (8.75 hours). Figure. 2(b) displays the allocation of slices for each class in UrbanSound8K, categorized by salience.

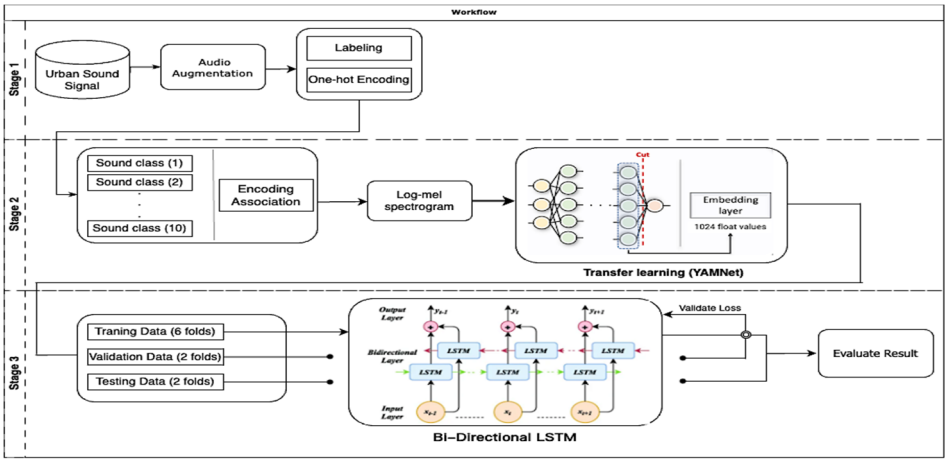

Figure 1: The workflow of the study

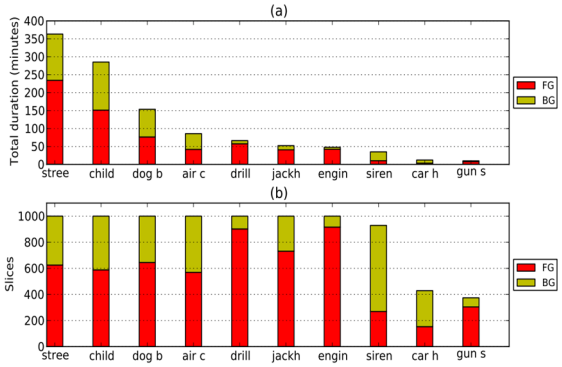

Figure 2 (a) Total duration of occurrences for each class within the UrbanSound dataset (b) the distribution of audio slices per class in UrbanSound8K, categorized into foreground (FG) and background (BG) components (Salamon, Jacoby, & Bello, 2014).

The UrbanSound8K dataset consists of audio files and a CSV file with metadata, including filenames, class designations, event start and end times, and fold numbers. Pre-structured into 10 folds for optimized cross-validation, it ensures consistent training and evaluation while preventing data leakage by keeping files from the same recording within the same fold. The dataset is partitioned into training, validation, and test sets: folds 1-6 for training, folds 7-8 for validation, and the remaining folds for testing.

2.2 Data Preprocessing

Each audio file in the UrbanSound8K dataset is labeled with one of 10 sound categories. To prepare these labels for the model, initial conversion of categorical labels into numerical format was achieved by label encoding. The numerical labels were then transformed into binary vectors via one-hot encoding. This step is crucial for multi-class classification, allowing the model to predict the probability of each class independently during training.

In this study, YAMNet, a pre-trained model developed by Google, was utilized for feature extraction. YAMNet is designed for general audio classification, capable of recognizing over 500 sound classes (detailed in Section 2.3). Leveraging YAMNet, high-level audio embeddings were obtained to be used as input features for the subsequent classification model. This process involved loading each audio file, processing it through YAMNet, and extracting a 1024-dimensional embedding vector from its penultimate layer. This embedding captures the essential features of the audio clip, making it suitable for machine learning models.

2.3 Model Architecture

2.3.1 Pre-Trained Model: YAMNet Structure

YAMNet processes input audio signals using a sequence of depthwise separable convolutional layers. Each convolutional block is followed by pooling layers, leading to fully connected layers and, ultimately, a softmax layer that outputs classification probabilities for 521 classes. The primary advantage of YAMNet is its ability to generate 1024-dimensional embedding vectors from raw audio inputs. These embeddings capture essential audio characteristics, providing rich representations crucial for subsequent classification tasks.

Leveraging these pre-trained embeddings reduces the need for extensive training data, thereby improving development efficiency. Additionally, YAMNet's architecture is optimized for resource-constrained environments, confirming optimal utilization in practical scenarios. Figure 3 illustrates the detailed architecture of YAMNet, showcasing the sequence of convolutional layers and dimensional transformations throughout the network.

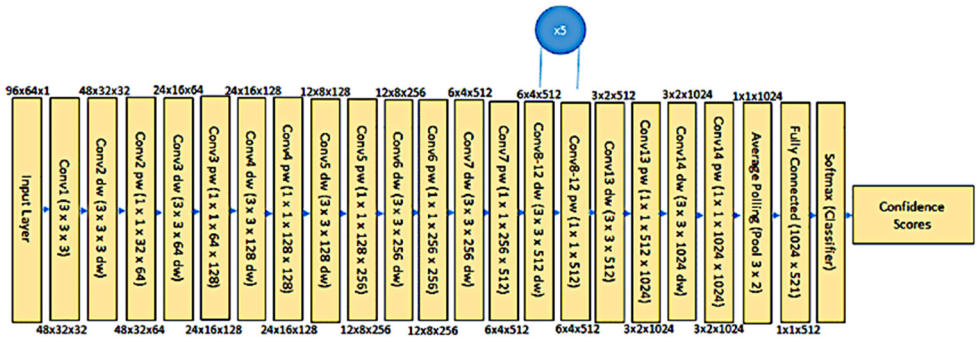

Figure 3. YAMNet Body Architecture (Tena, Claria, & Solsona, 2022). Conv: Convolution. dw: Depthwise. pw: Pointwise.

M YAMNet is efficient, making it suitable for implementation in settings with limited computational resources. This efficiency adds value to the urban sound classification pipeline, serving as a strong foundation for further processing stages.

2.3.2 Bi- Directional LSTM

The Bi-LSTM network is a crucial component in the proposed model architecture. Unlike traditional LSTMs, which process input sequences in a single direction (typically from past to future), BiLSTMs handle data in both directions. This flexibility allows the network to effectively capture both forward and bacward contexts, which is especially advantageous for tasks that include sequential input, such as audio categorization. In urban sound classification, the BiLSTM layer boosts the capability of the model to understand temporal dependencies in sound sequences by analyzing the entire sequence simultaneously. For example, BiLSTM can process the start, middle, and end of an audio clip simultaneously, allowing the recognition of patterns that a unidirectional LSTM might miss.

The BiLSTM layer in this model contains 128 units in each direction, allowing the model to acquire information from both preceding and following audio frames simultaneously. This dual processing enhances SED accuracy by integrating information from different parts of the audio clip to make more informed predictions. Integrating the Bi-LSTM layer into the model ensures that the temporal characteristics of the audio data are fully leveraged, leading to more accurate urban sound classification.

By integrating the Bi-LSTM layer into the model, the temporal characteristics of the audio data are fully leveraged, leading to more accurate classification of urban sounds. Figure.4 illustrates the structure of the Bi-LSTM network used in this study, showing how the forward and backward LSTMs process the input

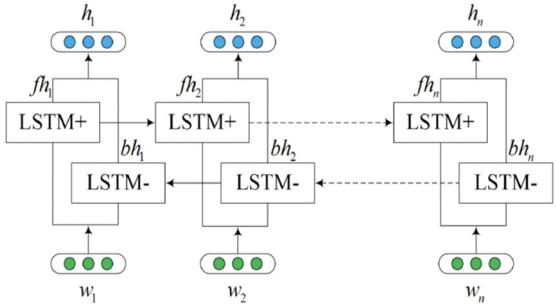

Figure 4: The architecture of Bi-LSTM (Xie, Chen, Gu, Liang, & Xu, 2019).

data, where $\{w_1, w_2, \ldots, w_n\}$ represent the word vector, and n is the length of one sentence. $\{fh_1, fh_2, \ldots, fh_n\}$ and $\{bh_1, bh_2, \ldots, bh_n\}$ represent the forward and backward hidden vectors, respectively, with h_n denoting the vector formed by connecting fh_n and bh_n.

2.4 Model Optimization

2.4.1 Loss Function:

In this study, Categorical Crossentropy was selected as the loss function due to its effectiveness in multi-class classification tasks. Categorical Crossentropy quantifies the discrepancy between the anticipated probability distribution and the actual data distribution, penalizing incorrect classifications more severely when the probability of the prediction deviates significantly from the actual label. This characteristic makes it particularly suitable for tasks like urban sound classification, where the model needs to distinguish between multiple sound classes.

The Categorical Crossentropy loss function promotes the model to provide high likelihoods for the accurate class selection while reducing the likelihoods for the incorrect classes. Categorical Crossentropy can be expressed by the following formula:

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \quad (1)$$

where: N represent the number of samples, C is the total number of classes. The true probability of the i-th sample belonging to the j-th class, denotes as y_{ij} (usually a one-hot encoded vector), $\hat{y}_{ij}$ is the predicted likelihood of the i-th sample being classified as the j-th class.

2.4.2 Optimizer:

The selection of the Adam (Adaptive Moment Estimation) optimizer for model training was based on its capacity to effectively manage extensive datasets and sparse gradients. The Adam optimizer integrates the advantageous features of two other widely used optimizers, namely AdaGrad and RMSProp. The algorithm adjusts the learning rate for each parameter separately, resulting in accelerated convergence and improved performance. In this work, the learning rate was initially established at 0.001, a widely employed figure that achieves a harmonious equilibrium between speed and stability throughout the training process. The default parameters of Adam ($\beta1$=0.9, $\beta2$=0.999, and ε=1e-7) were used, as they have been shown to work well across a wide range of tasks.

3 RESULTS AND DISCUSSION

3.1 Confusion Matrix

From the matrix shown in Figure 5, it is evident that certain classes, such as "jackhammer" and "engine idling," were occasionally misclassified, likely due to their similar acoustic profiles. This observation implies that the model could be improved by additional refinement strategies, possibly through enhanced feature engineering or data augmentation techniques, to better differentiate between these closely related sounds.

3.2 ROC Curves

Figure 6 displays Receiver Operating Characteristic (ROC) curves that offer a systematic assessment of the model's performance at various classification thresholds. For each sound class, each curve depicts the balance between the true positive rate and the false positive rate. The area under the receiver operating characteristic curve (AUC) is a crucial performance measure, where values approaching 1.0 indicate improved model performance.

In this study, the ROC curves obtained in this work have strong AUC values for all classes, with the "gun_shot" class achieving the AUC of 1.0, indicating exceptional accuracy. The "engine_idling" and "car_horn" classes also exhibit strong performance, with AUC values exceeding 0.99. However, classes such as "air_conditioner" and "dog_bark" present lower AUC values, suggesting that the model faces challenges in distinguishing these sounds under certain conditions. Overall, the ROC analysis underscores the model's high effectiveness in classifying urban sounds, especially for classes with distinct acoustic signatures.

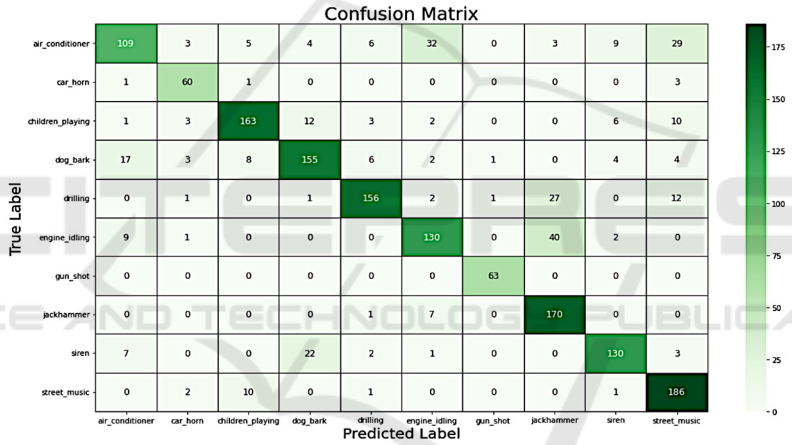

Figure 5: The confusion matrix for the training outcome

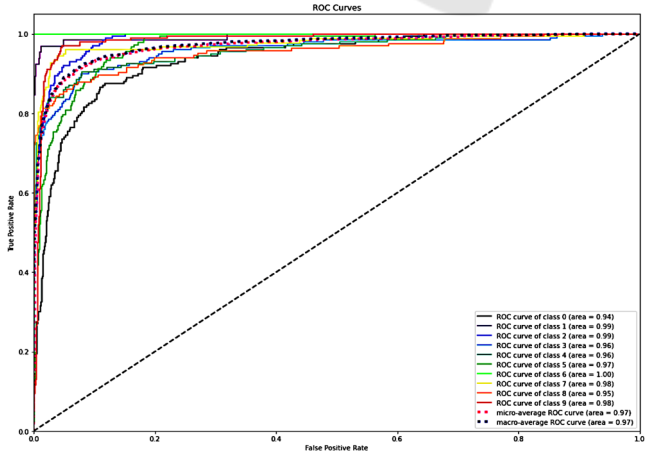

Figure 6: The ROC curves of training result

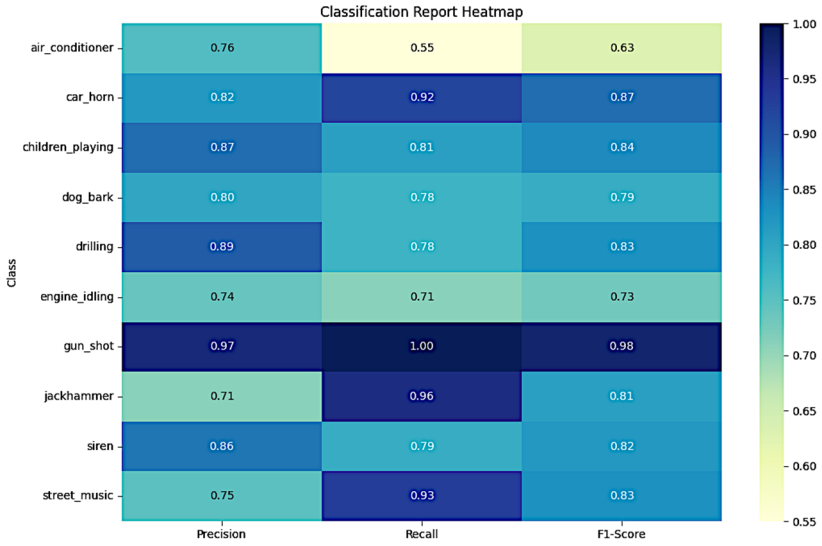

Figure 7: The precision, recall, F1-score of 10 tested classes

3.3 Performance Matrix

Figure 7 and Table 1 provide a summary of the results obtained from evaluating the performance of the proposed model on the UrbanSound8K dataset using various key criterion. A comprehensive analysis of the model's performance across the 10 sound classes is presented in Figure 7, including precision, recall, and F1-score. Notably, the model achieved a high level of accuracy in categories such as "gun_shot" (0.97) and "car_horn" (0.82), reflecting a strong ability to correctly identify these sounds when they are present. However, the model exhibited lower recall for classes like "air_conditioner" (0.55) and "dog_bark" (0.78), indicating challenges in detecting all instances of these sounds.

The YAMNet achieved an overall accuracy of 0.80, as shown in Table 1, the associated F1-score of 0.796 indicating a well-balanced performance in terms of precision and recall. Furthermore, the Matthews Correlation Coefficient (MCC) was 0.776, underscoring the model's capacity to produce effective predictions that exhibit a robust association with the true labels. Collectively, these metrics highlight the model's robustness in urban sound classification, while also indicating areas for potential improvement, particularly for classes with overlapping acoustic features.

Table 1: Aggregate Performance Metrics.

Metric	Quantity
Accuracy	0.79976
Precision	0.80492
Recall	0.79976
F1-Score	0.79628
Matthews Correlation Coefficient (MCC)	0.77688

4 CONCLUSION

In this study, an urban sound classification model was developed utilizing transfer learning, leveraging the UrbanSound8K dataset. By extracting high-level audio embeddings from YAMNet and integrating them with a Bidirectional LSTM network, a robust framework for urban sound classification was achieved. The model demonstrated strong performance, evidenced by a high Matthews Correlation Coefficient (MCC), indicating the statistical reliability of the model's predictions and capable of generalizing to unseen data. Despite these promising results, the model's relatively simple architecture suggests significant potential for further enhancement. Future improvements could focus on incorporating more advanced and complicated

architectures, such as transformer-based models, which have shown promise in handling complex audio signals. Additionally, exploring data augmentation techniques or fine-tuning YAMNet on domain-specific urban sound datasets could further enhance the model's capacity of distinguishing between classes that are acoustically quite similar.

REFERENCES

Nogueira, A. F. R., Oliveira, H. S., Machado, J. J. M., & Tavares, J. M. R. S. (2022). Sound classification and processing of urban environments: A systematic literature review. Sensors, 22(22), 8608. https://doi.org/10.3390/s22228608

Salamon, J., & Bello, J. P. (2015). Unsupervised feature learning for urban sound classification. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 171-175). IEEE. https://doi.org/10.1109/ICASSP.2015.7177954

Heittola, T., Mesaros, A., Eronen, A. J., & Virtanen, T. (2013). Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1), 1-15. https://doi.org/10.1186/1687-4722-2013-1

Cakir, E., Heittola, T., Huttunen, H., & Virtanen, T. (2015). Polyphonic sound event detection using multi-label deep neural networks. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE. https://doi.org/10.1109/IJCNN.2015.7280518

Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for polyphonic sound event detection. Applied Sciences, 6(6), 162. https://doi.org/10.3390/app6060162

Parascandolo, G., Huttunen, H., & Virtanen, T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. arXiv preprint arXiv:1604.00861. https://doi.org/10.48550/arXiv.1604.00861

Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(6), 1291-1303. https://doi.org/10.1109/TASLP.2017.2690575

Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2017). Large-scale weakly supervised audio classification using gated convolutional neural network. arXiv preprint arXiv:1705.02304. https://doi.org/10.48550/arXiv.1705.02304

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. https://doi.org/10.1109/JSTSP.2018.2885636

Turpault, N., Serizel, R., Shah, A., & Salamon, J. (2019). Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019) (pp. 253-257).http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Turpault_42.pdf

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

Tena, A., Claria, F., & Solsona, F. (2022). Automated detection of COVID-19 cough [Image of YAMNet Body Architecture]. Biomedical Signal Processing and Control, 71, 103175. https://doi.org/10.1016/j.bspc.2021.103175

Xie, J., Chen, B., Gu, X., Liang, F., & Xu, X. (2019). Self-attention-based BiLSTM model for short text fine-grained sentiment classification. IEEE Access, 7, 180558-180570.