
DEEP LEARNING NETWORK FOR AUDIO-BASED ENVIRONMENTAL SOUND RECOGNITION USING YAMNET AND DENSE SELF-ATTENTION MECHANISM

Shehu Mohammed Yusuf^{1,*}, Kabiru Muhammed Nasiru¹

¹Ahmadu Bello University

Email: smyusuf@abu.edu.ng, nasirumuhammedkabiru@gmail.com

*Corresponding author: Shehu Mohammed Yusuf

DOI: <https://doi.org/10.67203/sp.ICESID.2026>

ABSTRACT

Environmental sound recognition has become increasingly important in applications ranging from smart surveillance to health monitoring and intelligent urban systems. Although recent deep learning models have improved the ability to classify complex acoustic events, many of them still struggle with redundant features in audio embeddings or rely on heavy architectures that limit deployment on resource-constrained devices. To address these challenges, we developed a lightweight framework that combines YAMNet's pre-trained audio embeddings with a simple one-dimensional self-attention mechanism. This design helps the model focus on the most informative parts of each audio segment while avoiding unnecessary computational overhead. The enhanced embeddings are then processed through a compact fully connected network for final classification. Experiments on two benchmark datasets, ESC-10 and BDLib, demonstrate the effectiveness of the proposed approach, achieving test accuracies of 96% and 97%, respectively. The model outperforms an Xception-based transfer learning baseline on ESC-10 by 2.7% while using roughly 45× fewer parameters and provides competitive performance to MobileNetV2 on BDLib with only a fraction of its complexity. Though, the model was developed by utilizing small datasets with few classes of sounds, the results obtained indicate that applying self-attention directly to pre-trained audio embeddings can significantly improve environmental sound classification and offers a solution for real-time acoustic monitoring in noisy environments.

Keywords: Environmental sound classification; Neural Networks; Self-Attention Mechanism; transfer learning

1. INTRODUCTION

Environmental sound classification (ESC) is increasingly becoming an integral component of intelligent systems, particularly in applications such as automated surveillance, health monitoring, smart homes, and smart cities. These intelligent systems rely on the ability to recognize and interpret everyday acoustic events such as; natural sounds like rain or ocean waves and critical events such as alarms, sirens, or human activities. Also, ESC has found applications in smart surveillance systems, health monitoring, and environmental sensing in urban areas (Piczak, 2015a). For instance, in a security system, detecting a glass-breaking sound could help raise an alarm even in the absence of video data (Hershey *et al.*, 2017).

Benchmark datasets such as ESC-10 and ESC-50 have played an important role in enabling systematic evaluation of ESC methods by providing curated collections of labeled environmental sounds (Mu *et al.*, 2021). Compared with structured signals such as speech and music, environmental sounds often exhibit highly variable and complex time–frequency patterns, which makes them more challenging to model reliably (Mkrtchian & Furlotov, 2022).

Early ESC approaches were largely based on hand-crafted features such as Mel-frequency cepstral coefficients (MFCC) combined with classical machine learning algorithms. Although these pipelines gave useful baselines, they typically struggled to capture richer temporal and spectral dependencies and often degraded in the presence of noise or overlapping events (Mushtaq & Su, 2020). The emergence of deep learning has significantly improved ESC performance. Convolutional neural networks (CNNs) operating on spectrograms or stacked acoustic features have been shown to outperform traditional models on ESC-10, ESC-50 and related corpora (Mun *et al.*, 2017).

Recent works have further refined these architectures using multi-branch (Latifi *et al.*, 2024) or pyramidal CNNs to better capture multi-scale time–frequency patterns (Demir *et al.*, 2020; Fang *et al.*,

2022). A parallel line of research has focused on pre-trained audio models trained on large-scale datasets such as AudioSet. PANNs, for instance, provide large-scale pre-trained audio neural networks that have demonstrated strong transfer performance across a variety of audio pattern recognition tasks (Kong *et al.*, 2020). Similarly, several studies have adopted YAMNet-based transfer learning for downstream audio tasks. The YAMNet-based transfer learning model, a 1024-dimensional embedding generator strategy, is derived from a MobileNet-like architecture trained on hundreds of sound event classes (Liu *et al.*, 2025). This strategy has been particularly attractive in resource-constrained scenarios such as embedded anomaly detection in smart cities, where pre-trained models are retrained on ESC-10, BDLlib, and related datasets and then deployed on Raspberry Pi or TPU platforms (Lamrini *et al.*, 2023).

Despite these advances, two key limitations remain. First, embeddings produced by general-purpose pre-trained models may contain redundant or weakly relevant dimensions when applied to more specific ESC tasks. Second, CNN-based transfer learning architectures typically introduce millions of additional parameters, which increases memory consumption and computational latency. Thus, model developed may not be ideal for edge deployment. For instance, CNN-based transfer learning using Xception has achieved strong performance on ESC-10 and ESC-50 but at the cost of relatively large model sizes (Zhou *et al.*, 2025). At the same time, work on lightweight ESC models, such as resource-adaptive CNNs (Binandeh Dehaghani *et al.*, 2025; Fang *et al.*, 2022), has highlighted the importance of carefully balancing accuracy with computational efficiency for practical deployment.

Motivated by these observations, this study explores a lightweight framework that applies a one-dimensional self-attention mechanism directly to YAMNet embeddings for environmental sound recognition. The central ideas is to refine the pre-trained embedding space by emphasizing the most informative dimensions before classification, rather than adding deeper convolutional or recurrent stacks. As detailed in the rest of this paper, the proposed model combines YAMNet-derived 1024-dimensional embeddings with a simple dense self-attention layer and a compact fully connected classifier, and is evaluated on ESC-10 and BDLlib. The results show that this approach can achieve high accuracy while using substantially fewer parameters than typical CNN-based transfer learning baselines, making it a promising option for real-time ESC on embedded devices.

II. RELATED WORKS

A. Environmental Sound Classification with Deep Learning

Early deep learning studies showed that CNNs operating on log-Mel spectrograms can substantially outperform classical machine learning methods on these datasets. Piczak demonstrated that relatively shallow CNNs already capture useful local time–frequency patterns for ESC-50, while subsequent work improved robustness further using deeper architectures and data augmentation (Piczak, 2015b). The work of Mushtaq and Su (2020) proposed a regularized deep CNN with data augmentation, reporting improved performance on ESC-10, ESC-50 and UrbanSound8K by combining appropriate feature extraction with network regularization. Other studies have explored alternative CNN topologies such as concatenated or pyramidal architectures to capture richer multi-scale representations of environmental sounds (Mu *et al.*, 2021).

At the same time, there has been a growing interest in models that can provide strong performance while remaining computationally efficient. The work of [8] introduced a resource-adaptive CNN (RACNN) that uses a lightweight convolutional module to reduce hardware requirements and speed up ESC, highlighting the importance of designing architectures explicitly for embedded deployment. Other efficient end-to-end architectures, such as AclNet (Huang & Alvarado Leanos, 2018), and SwishNet (Hussain & Haque, 2018), have explored 1D CNNs and compressed model designs to achieve high accuracy at significantly reduced complexity. These works collectively show that careful architectural design can strike a useful trade-off between accuracy and resource usage.

B. Pre-trained Audio Models and YAMNet-based Transfer Learning

Large-scale pre-trained audio models have become increasingly influential in ESC and related tasks. PANNs, trained on AudioSet, provide pre-trained CNN backbones that can be fine-tuned or used as feature extractors for various audio tagging, acoustic scene classification, and sound event detection problems (Kong *et al.*, 2020). Building on this idea, several recent studies have adopted YAMNet as a

generic feature extractor. In their work, Liu *et al.* (2025) used YAMNet-based transfer learning to build a compact noise classifier. Their findings showed that its 1024-dimensional embeddings can be effectively reused in downstream tasks with relatively small additional networks. In the work of Wu (2025), a comprehensive approach for urban sound detection was developed using YAMNet obtain high-level embeddings and bidirectional LSTM as classification modules. The work of Mohino-Herranz *et al.* (2025) showed that the use of pretrained YAMNet for sound event classification on edge devices reduce training requirements and enable efficient deployment.

Lamrini *et al.* (2023) evaluated the performance of a pre-trained CNN for environmental sound recognition on multiple platforms, including laptops, Raspberry Pi, and Coral TPU, using ESC-10, BDLib, and UrbanSound datasets (Salamon *et al.*, 2014). Their results showed that transfer learning with pre-trained audio networks can achieve up to 100% accuracy on BDLib while remaining feasible on embedded hardware. However, these systems often rely on relatively heavy CNN backbones with millions of parameters. Although pre-training reduces the amount of labeled data required, the resulting models may still be over-parameterized for low-power deployment and may not explicitly address redundancy within the pre-trained embedding space.

C. Attention Mechanisms for Environmental Sound Classification

Attention mechanisms have been widely adopted to improve the modeling of complex acoustic patterns. In their work, Mu *et al.* (2021) proposed a temporal–frequency attention-based CNN (TFCNN) that applies attention over log-Mel spectrograms to focus on informative time–frequency regions. The work of Qiao *et al.* (2021) compared sub-spectrogram segmentation with temporal–frequency attention mechanisms and showed that attention mechanisms can provide performance gains by adaptively weighting different spectrogram regions. More recently, attention has been integrated into residual architectures and multi-branch CNNs for ESC (Zhong *et al.*, 2025). This further demonstrate its usefulness in refining spectro-temporal representations.

Beyond ESC, self-attention and multi-head attention have been explored in broader audio tasks. The work of Dong *et al.* (2023), applied multi-head attention to acoustic scene analysis and showed that attention over frame-level representations can yield competitive performance with only recording-level supervision. In the work of Wang *et al.* (2020), self-attention mechanism was assessed on learned features for joint sound event localization and detection. Findings demonstrated that self-attention can effectively replace or complement recurrent layers in CRNN-based architectures. Sudarsanam *et al.* (2021) developed an adaptive memory-controlled self-attention mechanism to improve polyphonic sound event detection by selectively emphasizing salient temporal patterns. These studies suggest that self-attention is a flexible tool for highlighting informative portions of audio representations; whether in time, frequency, or latent feature space.

Most attention-based ESC models, however, operate on spectrograms or intermediate CNN feature maps, which requires additional preprocessing and often increases model complexity. In contrast, the present work investigates a simpler alternative by applying a one-dimensional self-attention mechanism directly to pre-trained YAMNet embeddings and compact neural network classifier. As shown in our experiments, this embedding-level attention can yield competitive or superior ESC performance while maintaining a very small parameter budget. Thus, making our approach attractive for resource-constrained deployments.

III. METHODOLOGY

This section describes the proposed system for environmental sound classification; including data preprocessing, feature extraction using YAMNet, the application of a self-attention mechanism, and the final classification using fully connected layer. Fig. 1 shows the steps involved in process of audio classification.

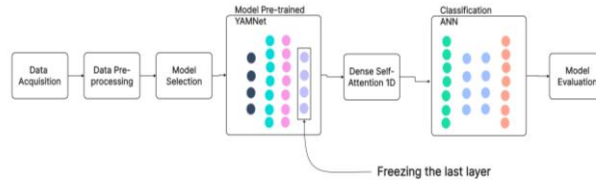


Figure 1: Proposed ESC framework

A. Dataset Aquisition and Preprocessing

Two benchmark datasets were used in this study: ESC-10 (Piczak, 2015a) and BDLib (Bountourakis *et al.*, 2015; Bountourakis *et al.*, 2019). The ESC-10 dataset was acquired from <https://github.com/karoldvl/ESC-50/archive/master.zip>. ESC-10 contains 400 audio samples across 10 environmental classes such as dog bark, clock ticking, and chainsaw. Each audio clip is 4 seconds long and sampled at 44.1 kHz. BDLib is a custom dataset collected for environmental monitoring applications. For this work, the BDLib dataset was acquired from <https://research.playcompass.com/files/BDLib-2.zip>. It includes sounds like rain, helicopter, sea waves, and sneezing, each stored as a .wav file in class-labelled folders.

To prepare the data, all audio files were first resampled to 16 kHz and converted to mono, matching the input requirements of YAMNet (Hershey *et al.*, 2017). A sliding window of 4 seconds was applied to each clip with a 50% overlap (Lamrini *et al.*, 2023). This resulted in more training samples from each audio file and ensured that important sound segments were not missed. Basic noise filtering was applied to remove silent frames. Fig. 2 shows the windowing process technique.

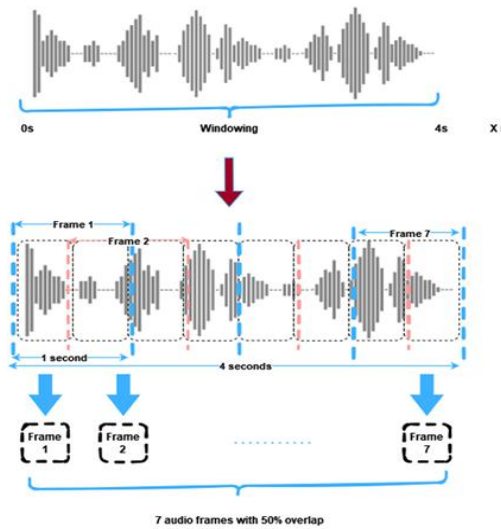


Figure 2: Windowing approach (Lamrini *et al.*, 2023)

B. Feature Extraction Using YAMNet

In this work, the YAMNet pretrained model was used purely as a feature extractor. Its classification head was removed, and the last fully-connected layer was frozen. The output of the frozen fully connected layer serves as input to the dense self-attention module of Figure 1. A 4-second audio window was adopted. Fig. 3 shows the adopted YAMNet architecture.

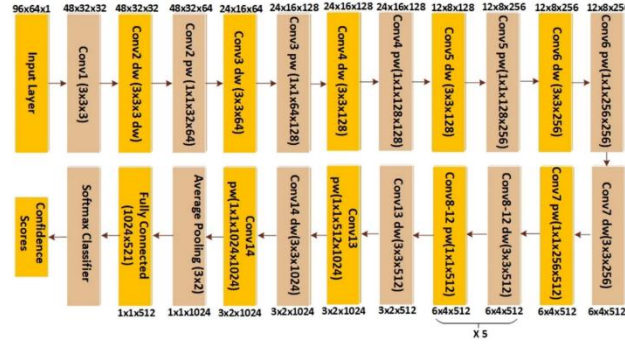


Figure 3: Adopted YAMNet architecture (Hershey *et al.*, 2017)

C. Self-Attention Mechanism Module

After feature extraction, a custom 1D self-attention layer was applied to the embeddings. This mechanism allows the model to assign different levels of importance to each dimension of the YAMNet output vector. The attention layer consists of a dense projection of the embedding followed by a SoftMax operation to compute attention scores. These scores are then multiplied element-wise with the original embeddings to produce an attention-weighted feature vector. Fig. 4 illustrate a single-head dense self-attention mechanism applied to 1D embeddings

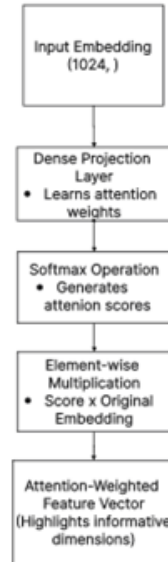


Figure 4: Single-headed dense self-attention module

This process enhances the most informative elements in the vector and suppresses those with little contribution to the final prediction. Unlike the temporal–frequency attention applied on spectrograms in earlier work (Kuzminykh *et al.*, 2020), this approach applies attention directly to a compact, pre-trained embedding. The goal is to reduce complexity while preserving performance.

D. Classification Module

The attention-enhanced embeddings generated by the single-headed dense self-attention module of Figure 4 are passed through a fully connected network consisting of two dense layers with 256 units each, using ReLU activation and dropout of 0.5 for regularization. Then, a final dense layer with 10 units and SoftMax activation was used to produce the probability distribution across the 10 sound classes. The summary of the architecture of the proposed model and parameter information used in each layer.

TABLE I. SUMMARY OF MODEL PARAMETERS

Type of Layer	Output Shape	Number of Parameters
Input_layer (InputLayer)	(None, 1024)	0
dense_self_attention1d (DenseSelfAttention1D)	(None, 1024)	131,329
dense_2 (Dense)	(None, 256)	262,400
dropout (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 256)	65,792
dropout_1 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 10)	2,570
Total parameters		462,091

E. Training Procedure

Two training scenarios were followed. First, the datasets were split into training (80%), validation (10%), and testing (10%) sets. The model was trained using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as the loss function. Training was conducted on the ESC-10 and BDLlib dataset for 10 and 20 epochs, respectively. Batch size was set to 32. The entire pipeline was implemented using TensorFlow and trained on a GPU-enabled environment.

Second, another training was conducted, using 5-fold cross validation (Zhou *et al.*, 2025) on the ESC-10 dataset, to allow for precise evaluation of the proposed architecture in handling small dataset.

Performance was measured using accuracy, precision, recall, and F1-score. The evaluation was done on the test set to ensure generalization. A confusion matrix was also used to visualize per-class prediction accuracy.

IV. RESULTS

The model was evaluated on two datasets: ESC-10 and BDLlib. Each dataset was divided into 80% training, 10% validation, and 10% testing splits.

A. ESC-10 Datasets Results

For the ESC-10 dataset, the model achieved a test accuracy of 96%, which shows that the proposed attention-based model can effectively classify environmental sounds using only the YAMNet embeddings. Table II shows the performance obtained based on the metrics for evaluation.

TABLE II. MODEL PERFORMANCE ON THE ESC-10 DATASET

Metric	Value
Accuracy	96%
Precision	98%
Recall	97%
f1-Score	97%

This result outperforms the recent transfer learning approach using the Xception CNN by Zhou *et al.* (2025); which reported 93.3% accuracy on the same ESC-10 dataset using mel-spectrogram-to-image conversion and a much larger model (~20.8 million parameters). In contrast, our lightweight architecture with only 462K trainable parameters achieves a 2.7% higher accuracy while being approximately 45 times more parameter-efficient. The confusion matrix for the ESC-10 dataset is presented in Fig. 5.

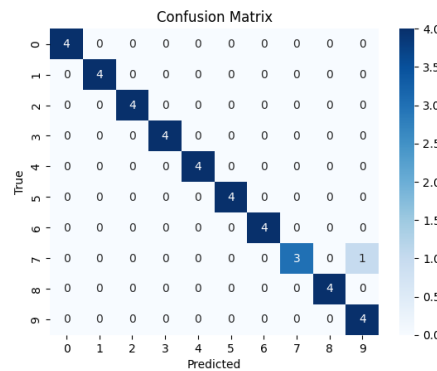


Figure 5: Developed model confusion matrix on the ESC-10 dataset

As seen in Figure 5, 9 classes were efficiently classified and only 1 test sample was misclassified; mostly in acoustically similar categories such as “clock ticking” and “crackling fire”.

Table III shows the cross-validation performance obtained based on the metrics for evaluation.

TABLE III. MODEL CROSS-VALIDATION PERFORMANCE ON THE ESC-10 DATASET

Metric	Value
Accuracy	95.40%
Precision	97.28%
Recall	96.29%
f1-Score	96.29%

Similarly, the cross-validation performance, in terms of accuracy, outperforms the recent transfer learning approach using the Xception CNN by Zhou *et al.* (2025) by 2.25%.

B. BDLib Dataset Results

The BDLib dataset yielded a test accuracy of 97% using the same model architecture. Table IV shows the performance obtained based on the metrics for evaluation.

TABLE IV. MODEL PERFORMANCE ON THE BDLIB DATASET

Metric	Value
Accuracy	97%
Precision	97%
Recall	97%
f1-Score	97%

Even though this dataset contains more real-world variations, the attention mechanism still performed well in learning meaningful embeddings from the YAMNet output. This result is competitive with the pre-trained MobileNetV2 transfer learning approach by Lamrini *et al.* (2023), which reported 100%, but required ~3.5 million parameters. Our lightweight design with only 462K trainable parameters achieves near-equivalent performance while being approximately 7.5 times more parameter-efficient. The confusion matrix for the BDLib dataset model is shown in Figure 6.

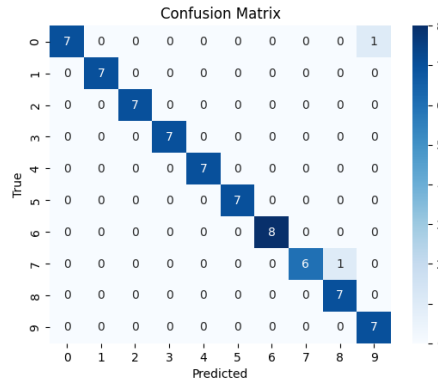


Figure 6: Developed model confusion matrix on the BDLib dataset

These results indicate strong generalization capability of the model on unseen, noisy samples.

V. DISCUSSION

The use of a self-attention mechanism on the YAMNet embedding played a key role in improving the classification performance. Compared to deeper CNN-based transfer learning models such as Xception used in the work of Zhou *et al.* (2025), which require over 20 million parameters to achieve 93.3% cross-validation accuracy on ESC-10 dataset, our YAMNet + self-attention model delivers superior performance with significantly lower computational overhead; proposed model shows 45 times smaller in size. Table V summarizes the comparison.

TABLE V. COMPARISON WITH BASELINE ON ESC-10 DATASET

Method	Accuracy	Parameters	Parameter Ratio of baseline to Proposed Model
Developed Model (YAMNet + Self-Attention) using 5-fold Cross-validation	95.4%	462K	~45
(Zhou <i>et al.</i> , 2025) (Xception)	93.3%	~20.8M	

On the custom BDLib dataset, the 97% accuracy further validates the model's generalization to varied environmental sounds, closely approaching the 100% achieved by Lamrini *et al.* (2023) with MobileNetV2-based transfer learning under a matching 80/10/10 split. Table VI summarises the comparison.

TABLE VI. COMPARISON WITH BASELINE ON BDLIB DATASET

Method	Accuracy	Parameters	Parameter Ratio of baseline to Proposed Model
Proposed Model (YAMNet + Self-Attention)	97%	462K	~7.5
Lamrini <i>et al.</i> (2023) (BDLib-based MobileNetV2)	100%	~3.5M	

Similarly, from Table VI, the proposed model shows over 7.5 times improvement over the model developed in Lamrini *et al.* (2023), in terms of model size. This parameter efficiency gain, combined with high accuracy, positions the developed method as a practical solution for real-time audio-based environmental monitoring in edge environments. The network was able to highlight critical dimensions of the embedding vector while suppressing noise, resulting in a more discriminative feature space. This is consistent with previous works which have shown that attention can improve model performance. Additionally, the compact structure of the network with only 462K parameters makes it suitable for deployment on resource-constrained devices such as mobile phones or embedded systems for smart surveillance and monitoring.

VI. CONCLUSION

This study presented a lightweight and effective deep learning architecture for environmental sound classification. The approach combines YAMNet embeddings with a custom 1D self-attention mechanism followed by fully connected layers. The model was trained and evaluated on two datasets; ESC-10 and BDLib. A 5-fold cross-validation accuracy of 95.4 was obtained for the ESC-10 based model. And a test accuracy of 97% was obtained for the BDLib-based model. Thus the use of a pre-trained feature extractor allowed the model to learn high-level representations of audio without requiring complex spectrogram-based processing. Additionally, the attention mechanism significantly improved the model's ability to focus on the most informative parts of the audio embedding, leading to better classification performance. The architecture also remains computationally efficient compared with the baseline models. Therefore it can be a more suitable candidate for real-world applications where speed and resource constraints are critical.

Future work may explore the use of hybrid attention mechanisms, domain adaptation techniques, or lightweight convolutional variants to further improve robustness across a wider range of sound environments. Also, there is the need to deploy the developed model on embedded platforms, like Raspberry Pi, to evaluate its inference latency against existing works, because parameter efficiency alone may not translate to real-time feasibility. Finally, there is the need to further subject the developed technique on large and balanced datasets with more categories of sounds; to evaluate its generalization capability.

ACKNOWLEDGMENT

This work is supported by TETFUND Nigeria under the grant: NRF-2023-SETI-AFS-00158

REFERENCES

- Binandeh Dehaghani, P., Pena, D., & Aguiar, A. P. (2025). Investigation of feature selection and pooling methods for environmental sound classification. *arXiv*.
- Bountourakis, V., Vrysis, L., Konstantoudakis, K., & Vryzas, N. (2019). An enhanced temporal feature integration method for environmental sound recognition. *Acoustics*, 1(2), 410–422.
- Bountourakis, V., Vrysis, L., & Papanikolaou, G. V. (2015). Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. In *Proceedings of the 10th Audio Mostly: A Conference on Interaction With Sound*. Association for Computing Machinery.
- Demir, F., Turkoglu, M., Aslan, M., & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. *Applied Acoustics*, 170, Article 107520.
- Dong, S., Xia, Z., Pan, X., & Yu, T. (2023). Environmental sound classification based on improved compact bilinear attention network. *Digital Signal Processing*, 141, Article 104170. <https://doi.org/10.1016/j.dsp.2023.104170>.
- Fang, Z., Yin, B., Du, Z., & Huang, X. (2022). Fast environmental sound classification based on resource adaptive convolutional neural network. *Scientific Reports*, 12, Article 6599.
- Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131–135). IEEE.
- Huang, J. J., & Alvarado Leanos, J. J. (2018). AclNet: Efficient end-to-end audio classification CNN. *arXiv*.
- Hussain, M. S., & Haque, M. A. (2018). SwishNet: A fast convolutional neural network for speech, music and noise classification and segmentation. *arXiv*.
- Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2880–2894.

- Kuzminykh, I., Shevchuk, D., Shiaeles, S., & Ghita, B. (2020). Audio interval retrieval using convolutional neural networks. In O. Galinina, S. Andreev, S. Balandin, & Y. Koucheryavy (Eds.), *Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 20th International Conference, NEW2AN 2020 and 13th Conference, ruSMART 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12525, pp. 229–240)*. Springer.
- Lamrini, M., Chkouri, M. Y., & Touhafi, A. (2023). Evaluating the performance of pre-trained convolutional neural network for audio classification on embedded systems for anomaly detection in smart cities. *Sensors*, 23(13), Article 6227.
- Latifi, S. A., Ghassemian, H., & Imani, M. (2024). Classification of heart sounds using multi-branch deep convolutional network and LSTM-CNN. *arXiv*.
- Liu, L., Xu, Q., Mao, S., Mu, J., Zhao, X., Song, W., & Cheng, P. (2025). YAMNet-based transfer learning for compact noise classification in urban and wireless systems. *EURASIP Journal on Wireless Communications and Networking*, 2025, Article 74.
- Mkrtchian, G., & Furlotov, Y. (2022). Classification of environmental sounds using neural networks. In *2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)* (pp. 1–4). IEEE.
- Mohino-Herranz, I., García-Gómez, J., Alonso-Díaz, S., Gallegos, J. G., Perez-Sanz, F. J., Aguilar-Ortega, M., & Gil-Pita, R. (2025). Implementing transfer learning for sound event classification using the ReaLISED audio database. *Measurement: Sensors*, 38, Article 101711.
- Mu, W., Yin, B., Huang, X., Xu, J., & Du, Z. (2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11, Article 21552.
- Mun, S., Park, S., Han, D. K., & Ko, H. (2017). Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)* (pp. 93–97).
- Mushtaq, Z., & Su, S.-F. (2020). Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167, Article 107389. DOI: 10.1016/j.apacoust.2020.107389.
- Piczak, K. J. (2015a). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia* (pp. 1015–1018). Association for Computing Machinery. <https://doi.org/10.1145/2733373.2806390>.
- Piczak, K. J. (2015b). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE.
- 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). Association for Computing Machinery.
- Qiao, T., Zhang, S., Cao, S., & Xu, S. (2021). High accurate environmental sound classification: Sub-spectrogram segmentation versus temporal-frequency attention mechanism. *Sensors*, 21(16), Article 5500.
- Sudarsanam, P. A., Politis, A., & Drossos, K. (2021). Assessment of self-attention on learned features for sound event localization and detection. In *Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021)* (pp. 100–104). DCASE.
- Wang, W., Wang, W., Sun, M., & Wang, C. (2020). Acoustic scene analysis with multi-head attention networks. In *Proceedings of Interspeech 2020* (pp. 1191–1195). ISCA.
- Wu, T. (2025). A comprehensive approach to urban sound detection with YAMNet and bi-directional LSTM. In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML)* (pp. 23–29). SciTePress.
- Zhong, X., Dong, S., Zhang, Y., Lu, W., & Jiang, L. (2025). Research on multi-branch residual connection spectrum image classification based on attention mechanism. *Scientific Reports*, 15, Article 25532.

Zhou, J., Lu, J., Wang, R., Ma, R., & Qin, Z. (2025). Environmental sound classification using convolutional neural network based transfer learning. In 2025 International Conference on Computing, Networking and Communications (ICNC) (pp. 12–17). IEEE.