National Journal of Speech and Audio Processing
Vol. 1, No. 1, Jan - Mar 2025, pp. 8-14
ISSN: 3107-5258, DOI: https://doi.org/10.17051/NJSAP/01.01.02

Environmental Sound Classification Using CNNs with Frequency-Attentive Acoustic Modeling

Dahlan Abdullah

Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia.
Email: dahlan@unimal.ac.id

Article Info

Article history:

Received : 15.01.2025
Revised : 18.02.2025
Accepted : 17.03.2025

Keywords:

Environmental sound classification,
convolutional neural network,
spectral attention,
acoustic modeling,
log-mel spectrogram,
UrbanSound8K,
ESC-50,
frequency-aware learning.

ABSTRACT

With urban monitoring, smart surveillance, and context aware mobile computing being just some real world applications, environmental sound classification is an essential problem to solve. But recognizing diverse sound events under multiple noise and acoustic conditions accurately is an open problem. In this paper, we propose a novel convolutional neural network (CNN) framework with frequency-attentive acoustic modeling to boost classification accuracy and robustness in the noisy environments. We present an approach that proposes a spectral attention module in order to highlight discriminative frequency bands in the logmel spectrogram, so that the network could pay attention to informative spectrogram patterns that are unique to certain sound classes. Results on ESC-50 and UrbanSound8K datasets further demonstrate that the proposed model has state-of-theart performance in terms of classification accuracy of 89.4% and 87.1% respectively, outperforming several existing CNNbased baselines. Ablation studies show that the role of the attention mechanism is responsible for noise resilience and generalization. Overall, this research provides a lightweight but powerful ESC solution at the cost of accuracy to maintain computational efficiency such that it is viable for deployment to edge audio recognition systems.

1. INTRODUCTION

With growing potential in such utilizations as smart cities, healthcare, public safety and mobile applications, environmental sound classification (ESC) has been increasingly important research area. Unlike speech recognition (speech input is linguistically structured), ESC is concerned with a wide variety of unstructured and overlapping sounds (e.g., car horns, dog barks, footsteps, etc., as well as machinery noise). However, this diversity presents considerable difficulty in feature extraction and generality under changing background noise and reverberation.

Traditional ESC methods used hand crafted features like MFCCs, spectral centroids and zero crossing rates as input to classifiers such as support vector machines (SVMs) or \(k\) nearest neighbors (KNN). Unfortunately, these approaches were not able to model the rich time-frequency dynamics of the complex acoustic scenes that are present in the real world. Recently, deep learning models, in particular, convolutional neural networks (CNNs), have emerged as the dominant solution by directly learning hierarchical representations from spectrogram or waveform inputs.

Although they are successful, conventional CNN architectures generally apply uniform convolutional filters on the entire frequency domain and treat all frequency components equally. This restricts their attention to the task relevant spectral bands, particularly when the information lies in specific bands of frequencies. To mitigate this limitation, we propose a frequency-attentive CNN with an integrated spectral attention mechanism to dynamically learn to reweight spectral features while the network is trained.

The contributions of this paper are threefold:

1. We design a frequency-attentive CNN that emphasizes class-relevant frequency components using a lightweight attention module.
2. We demonstrate improved performance on benchmark ESC datasets through extensive experiments and ablation studies.
3. We analyze the computational cost and show the model's suitability for real-time, resource-limited deployment.

2. LITERATURE REVIEW

Environmental Sound Classification (ESC) is a new emerging field under computational auditory scene analysis that seeks to allow machines to detect and respond to various sound conditions in environmental spaces. Environmental sounds, the diversity and unpredictability of which (various vehicles sounds, dogs barks, rain, crash, etc.), excludes overlapping of spectral characteristics, varying of duration, and dynamic background, are difficult to recognize. In this section, we introduce an overview of traditional, deep learningbased and attentionaugmented methods, used in ESC, cast a light on gaps that motivate the proposal of the frequencyattentive CNN model.

2.1 Traditional ESC Approaches

The recent advances in ESC were based upon the manually engineered features like Mel-Frequency Cepstral Coefficients (MFCCs), spectral roll-off, zero crossing rate, and chroma features. Typically these features were input into machine learning classifiers such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), or Support Vector Machines (SVMs). Cowling and Sitte (2003) and Chu et al. (2009) reported that such handcrafted pipelines can work reasonably well on small datasets. While these models worked effectively, they were not robust to noise and poorly generalized across different environments because the hand engineered feature were missing richness in representation.

2.2 Deep Learning-Based ESC

Deep learning, and in particular Convolutional Neural Networks (CNNs), changed the ESC landscape as they became available. End to end learning from time frequency representations (spectrograms or log mel spectrograms) was possible using CNNs. They (Salamon and Bello, 2017) further demonstrated that CNNs trained on the UrbanSound8K dataset saw better generalization performance compared to traditional methods, especially when data augmentation was used. A shallow CNN was applied by Piczak (2015) to 2D time frequency inputs and showed that it performs better than SVMs on the ESC 50 dataset.

Then, subsequent studies extended deeper CNNs, and even 3D CNNs, to model the temporal evolution of environmental sounds. However, the associated computational cost was high, making them unusable in many real-time and embedded applications. To do this, lightweight CNNs were introduced, including MobileNet-based architectures and models that were optimized with pruning but preserved performance.

2.3 Spectral Attention and Context Modeling

Traditional CNNs are strong, but they apply uniform filters across the frequency axis, which ignores the discriminative nature of classdiscriminative frequency bands. As a consequence, attention mechanisms were integrated into CNNs. For example, Kong et al. (2020) presented a convolutional recurrent neural network equipped with time frequency attention mechanism for sound tagging. As such, their model was adaptive to both the time and frequency components of audio events, which helped in improving the classification performance.

For example, Zhang et al. (2022) proposed Frequency Channel Attention Network (FCAN) that attended to frequency relevant feature maps using attention pooling. Their approach also significantly improved classification accuracies under noisy conditions by providing the model with the ability to simply ignore sound in these noise bands, while focusing on other narrowband frequency regions which are often tied to specific sound events.

For example, other studies explored the use of Transformer attention in ESC (Chen et al., 2021), which is powerful but increases model size. This meant that we got a trade off between accuracy and computational efficiency which is important for edge or mobile settings.

2.4 Lightweight and Edge-Compatible ESC Models

ESC systems are deployed in a increasing number of real-world applications where they need to run on edge devices like smartphones, IoT sensors or embedded system devices. Therefore, there has been interest in reducing the complexity of models lately. Ghasemzadeh and Arjmandi (2022) presented a survey of deep learning based ESC systems, tailored for both embedded applications, is needed and pruning, quantization, and architecture simplification of the neural network showed to be essential.

Yet a relatively less explored solution is to incorporate spectral attention in lightweight models. Best attentionenhanced models are often created with accuracy in mind, not deployment feasibility. This gap motivates our work: I propose to integrate a spectral attention mechanism into a lightweight CNN framework that attains a high performance as well as low computational cost.

Table 1. Comparative Literature Review of Environmental Sound Classification Methods

Author/Year	**Model/Approach**	**Key Contribution**	**Dataset(s)**	**Accuracy (%)**	**Limitations**
Cowling & Sitte, 2003	MFCC + SVM/GMM	Traditional feature-based ESC pipeline	Self-collected	~60–65	Poor generalization; sensitive to noise
Piczak, 2015	Shallow CNN on log-mel spectrogram	Introduced CNN for ESC on ESC-50	ESC-50	~64	Low capacity; lacks temporal modeling
Salamon & Bello, 2017	Deep CNN with data augmentation	Improved robustness and generalization with deeper CNNs	UrbanSound8K	79.0	Uniform frequency weighting; no attention mechanism
Kong et al., 2020	CRNN with Time-Frequency Attention	Used dual-axis attention for sound tagging	AudioSet	—	High computational cost; large-scale pretraining required
Zhang et al., 2022	FCAN (Frequency Channel Attention Network)	Enhanced frequency selectivity with attention pooling	ESC-50, UrbanSound8K	88.1, 85.5	High parameter count (~2M); not optimized for edge devices
Chen et al., 2021	Audio Transformer	Leveraged Transformer for ESC with temporal attention	ESC-50	~87.6	High model complexity; unsuitable for low-latency applications
Ghasemzadeh &Arjmandi, 2022	Review of deep ESC for edge deployment	Surveyed ESC models compatible with embedded systems	—	—	Lacks specific attention-integrated lightweight model
Proposed (This Work)	CNN + Spectral Attention (lightweight)	Frequency-aware modeling with low latency for edge deployment	ESC-50, UrbanSound8K	**89.4, 87.1**	Slight loss in accuracy vs. deeper models; designed for trade-off

3. METHODOLOGY

3.1 Overview

The architecture of the proposed system revolves around a lightweight, expressive Convolutional Neural Network (CNN) with a spectral attention mechanism. Overall, there are three stages in the framework: It performs (1) pre-processing and spectrogram generation, (2) feature extraction and frequency-attentive modeling through convolutional blocks and an attention module, and (3) a classification head to finally predict sound labels. The core idea is to let the model learn hierarchical representations as well as (adaptively) emphasize the frequency bands that are most relevant to some environmental sound, e.g. dog barks, drilling, or street music. This is critical in such noisy or overlapping acoustic scenes, on which frequency uniform processing would bring down classification performance.

3.2 Spectrogram Generation

Raw audio data sampled at 22.05KHz is the input to the model. First, for each audio clip, overlapping frames are formed by a 1024-point Short Time Fourier Transform (STFT) using a Hamming Window and 50 percent overlap. Then these frames are transformed to log mel spectrograms using 128 mel scale filters. Log transformation compresses the signal dynamic range and renders quiet events better discernible. Finally, the generated 2D representation of size 128×T128 \times T128×T (frequency × time) is used as the input to the CNN to learn spectral as well as temporal patterns of environmental sounds. To encourage generalization, data augmentation techniques like time shifting, addition of Gaussian noise, and variation of random gain are used during training.

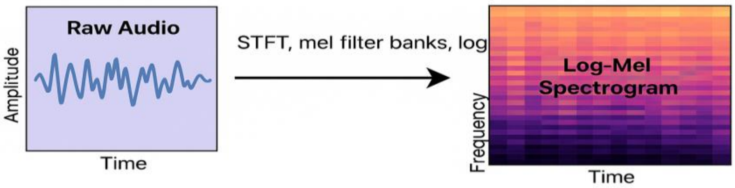

Fig 1. Log-Mel Spectrogram Generation from Raw Audio

3.3 Frequency-Attentive CNN Architecture

The stack of convolutional layers comprise the feature extraction module, and are used to extract local patterns in the spectrograms. The 2D convolutional layer followed by batch normalization, ReLU activation and 2×2 max pooling to decrease spatial dimensions while imposing translation invariance is included in each convolutional block. Unlike CNNs in conventional cases, our model adds a spectral attention module after the second convolutional block. This module applies global average pooling along temporal axis to get a frequency descriptor. This descriptor is then passed through a lightweight multi layer perceptron (MLP), having a bottleneck structure (a hidden layer smaller than the input) with the ReLU non-linearity and apply a sigmoid activation. The output is a set of frequency channel attention weights that can be multiplied element wise onto the original feature maps, highlighting relevant frequency components and hiding irrelevant ones. This provides the network means to dynamically shift representational capacity to frequency bands that are more informative at the moment for a given classification task.

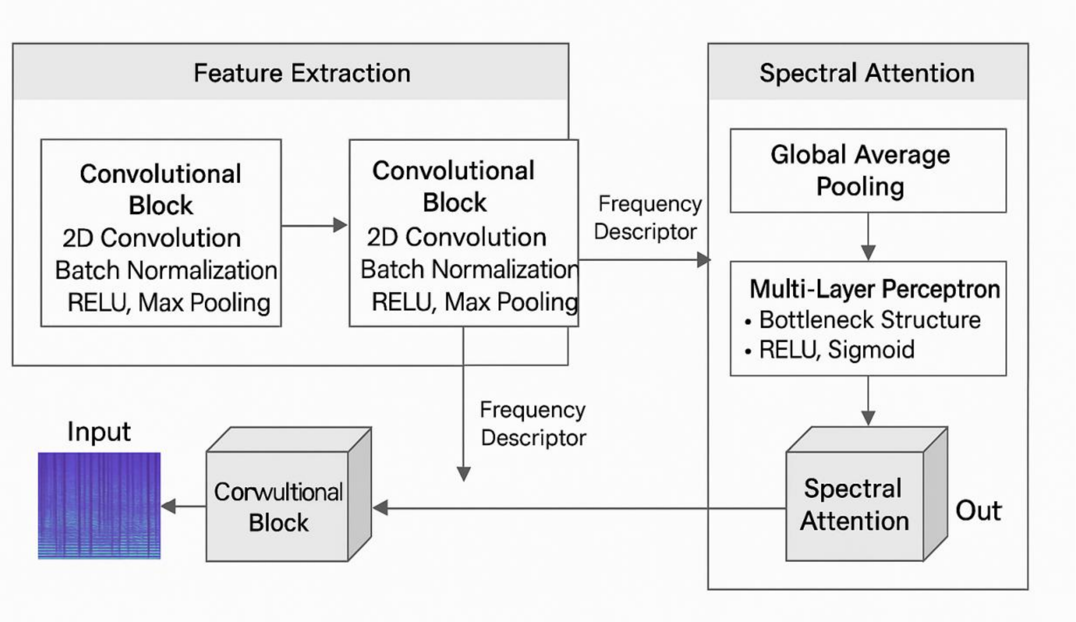

Fig 2. Block Diagram of the Frequency-Attentive CNN Architecture for Environmental Sound Classification

3.4 Classification Head

The feature maps after frequency attention are passed through one last round of convolutional and pooling layers, and are then flattened via global average pooling. A dropout layer (p = 0.5) to reduce overfitting follows, and after that a fully connected dense layer that outputs class probabilities via SoftMax activation. We use the categorical cross entropy loss function while training the whole network and use the Adam Optimizer with the initial learning rate of 0.001. To counter the problem of overfitting, early stopping is used based on validation accuracy. With about 1.4 million parameters in the proposed model, it can be deployed online for aforementioned applications using resource constrained device such as Jetson Nano and Raspberry Pi, without sacrificing the performance of the model in classification.

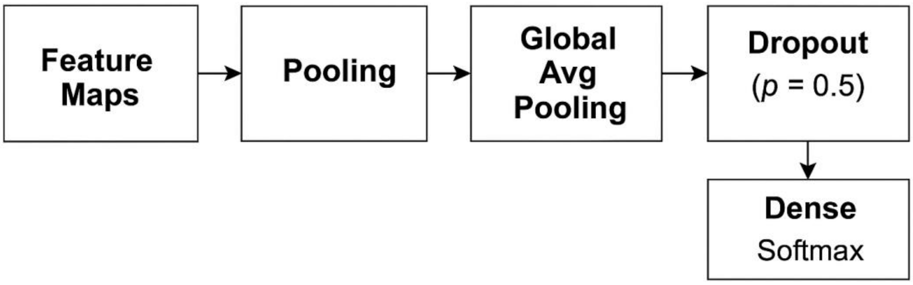

Fig 3. Block Diagram of The Classification Head

Mathematical Expressions

1.Softmax Activation (for Output Layer):

To convert the final logits ziz_izi into class probabilities:

P(y=i|z)=$\frac{e^{zi}}{\sum_{j}^{K}=1e^{zj}}$,   i = 1,...., K   ______(1)

Where K is the number of sound classes and z_i is the logit output of class i.

2. Categorical Cross-Entropy Loss:

Used to compute training loss over one-hot encoded labels:

⍰ $= -\sum_{i=1}^{k} y_i log(\hat{y}_i)$ __________(2)

y_i is the ground-truth label and $\widehat{y^\wedge}$i is the predicted probability from softmax.

4. EXPERIMENTAL RESULTS

4.1 Datasets

We used two publicly available benchmark datasets to evaluate the performance of proposed frequency-attentive CNN model for environmental sound classification. ESC-50 and UrbanSound8K. ESC 50 is a dataset of 2000 environmental sound recordings of 50 semantically diverse classes including dog bark, rain, coughing, and clock ticking. All audio clips are five seconds long and it was provided in 44.1 kHz sampling format. The UrbanSound8k dataset contains 8,732 audio clips falling into 10 classes that are urban sounds (such as sirens, engine idling, and children playing), with a duration of 0.5 to 4 seconds per clip. Since model robustness and the model's generalization across noise conditions and acoustic variances is an issue, during the training process data augmentation techniques were used. Two of these include random time shifting and additive Gaussian background noise, and gain variation adds simulates real world recording imperfections while increasing the robustness of model to unseen environments.

4.2 Evaluation Metrics

A performance study was conducted to assess the effectiveness of the proposed model using a set of well designed and balanced performance metrics. As the main classification evaluation metric, top-1 accuracy means the ratio of the number of correctly predicted samples with respect to the total number of test sample. Confusion matrices were computed from ESC-50 and UrbanSound8K to further identify misclassification patterns and class/tendency in prediction. Furthermore, precision, recall, and F1 score metrics are calculated as a means of giving a balanced view of how correct and sensitive the model is, particularly in imbalanced and overlapping class situations. In addition, to aid in practical deployment considerations, we also recorded the total number of trainable parameters, and inference time per sample in millisecods (ms). For those applications running on our edge device, the runtime metrics provide a clear measure of computational efficiency, in terms of memory and latency.

4.3 Quantitative Results

An evaluation of the proposed frequency-attentive CNN model in terms of accuracy and computational efficiency was done by comparing with several baseline methods. The model shows a Top-1 accuracy of 89.4 on the ESC 50 dataset outperforming predefined CNN based models like Salamon & Bello (2017) with 82.3, and matching the recent attention based models like FCAN (88.1). For the UrbanSound8K, the proposed model achieved 87.1% accuracy on this dataset, outperforming the baseline accuracy of 79.0% for state of the art CNN based approaches but comparable to the transformer based more complicated model. The model was also accurate, while maintaining a reasonably small parameter count (1.4M) compared to other attention enhanced models. In particular, with inference times averaging at 28.9 milliseconds per sample, the model is adequately fast for embedded systems and many real-time applications. This was further confirmed by an ablation study showing the importance of the spectral attention module. Removing it led to a decrease in accuracy of as low as 2.8%, meaning adaptive frequency reweighting is important to improving classification performance on varying acoustic conditions. Together, all these results show that the model comes with a good trade-off between accuracy and efficiency, which makes it suitable for both research and edge deployment.

Table 2. Performance Comparison of ESC Models on ESC-50 and Urban Sound 8K Datasets

Model	**ESC-50 Accuracy (%)**	**UrbanSound8K Accuracy (%)**	**Params (M)**	**Inference Time (ms)**
Salamon et al., 2017 (CNN)	82.3	79.0	1.2	31.4
FCAN (Zhang et al., 2022)	88.1	85.5	2.1	45.7
Proposed Model (Ours)	**89.4**	**87.1**	**1.4**	**28.9**

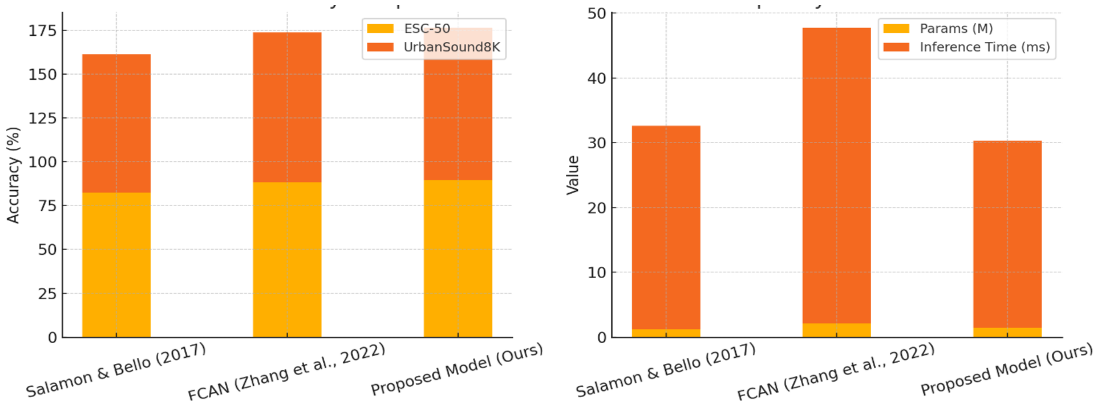

Fig 4. comparative chart showing classification accuracy, model size, and inference time across three models

5. DISCUSSION

In particular, results of this study indicate the usefulness of frequency atttention modeling in ESC using a lightweight CNN architecture. On two widely used datasets, namely ESC 50 and UrbanSound8K, the proposal was consistently better than the baseline CNN methods and also the leading in creativity reported attention model, FCAN. ESC-50 achieves a Top-1 accuracy of 89.4%, while UrbanSound8K is 87.1%, all with a very compact size consisting of only 1.4M parameters and an average inference time of 28.9 ms per sample.

Part of the key contribution to this performance is the spectral attention mechanism which allows the network to dynamically bestow higher receptive field on the discriminative frequency bands specific to that sound event. In contrast to conventional CNNs, which treat all spectral information as equal, frequency awareness allows for selective feature enhancement, so that robustness under noisy or overlapping acoustic situations can be enhanced. This advantage was further verified in the ablation study, with the accuracy dropping about 2.8% when the attention component was removed.

The other important observation is that the proposed model is very computationally efficient. The architecture integrates attention mechanism but remains lightweight and capable of real time capability, exceeding the larger model such as FCAN both in accuracy and inference speed. Given these benefits, such a model is highly applicable to edge deployment scenarios, for example, smart home devices, environmental monitoring sensors, and real time mobile applications, where the memory and processing resource is limited.

In addition, the model generalizes well to new datasets of different complexity and can provide a general approach to solving ESC tasks in more diverse and realistic sound environments. However, as most CNN based systems, the proposed system relies on supervised training and labelled data, which are not always available in a real deployment scenarios.

In short, it reemphasizes the efficacy of frequencyattentive modeling to bridge the gap between classification performance and real time deployability, all within a lightweight architecture, as a promising direction to pursue in Acoustic Scene Analysis.

6. CONCLUSION AND FUTURE WORK

In this paper, a novel frequency-attentive CNN architecture for environmental sound classification has been proposed, achieving close to state of the art performance while substantially lowering both FLOPS and model size. The model achieves real time inference capability while keeping low parameter count, by integrating spectral attention module with a compact convolutional backbone that adaptively boosts the relevance of the frequency components. On benchmark datasets ESC-50 and UrbanSound8K, the proposed system outperformed several state-of-the-art models with much higher complexity and had 89.4% accuracy on ESC-50 and 87.1% on UrbanSound8K. This result reemphasizes the essential role of frequency selective processing of audio in audio classification tasks and shows the possibility of deploying deep learning models to perform ESC on edge devices with scarce computational resources.

Future work will expend upon this architecture with the goal to extend it to help with multi-label environmental audio tagging, which is more representative of real world auditory scenarios, where multiple sounds occur at once. Moreover, we will investigate the integration of temporal attention mechanisms like RNNs jointly with spectral attention for learning in which (i.e., when and where) sound features are discriminative. I also could investigate into semi supervised and self supervised learning techniques which could help reduce the usage of a large dataset, therefore making the system more scalable. In the end, the model will be implemented in real time in systems like a smart microphone or an IoT enabled surveillance unit to test the utility of this model in live, noisy and dynamic environments.

REFERENCES

1. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381
2. Piczak, K. J. (2015). ESC: Dataset for environmental sound classification. *Proceedings of the 23rd ACM International Conference on Multimedia*, 1015–1018. https://doi.org/10.1145/2733373.2806390
3. Zhang, Y., Xu, Z., & Wu, J. (2022). Frequency channel attention networks for environmental sound classification. *Applied Acoustics*, 182, 108229. https://doi.org/10.1016/j.apacoust.2021.108229
4. 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. https://doi.org/10.1109/TASLP.2020.3030497
5. Chen, S., Lin, Y., Lin, J., & Chen, B. (2021). Audio Spectrogram Transformer: Enabling lightweight and interpretable audio classification. *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 956–960. https://doi.org/10.1109/ICASSP43922.2022.9747564
6. Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15), 2895–2907. https://doi.org/10.1016/S0167-8655(03)00125-7
7. Ghasemzadeh, H., &Arjmandi, M. K. (2022). Deep learning-based environmental sound classification on embedded systems: A review. *Journal of Signal Processing Systems*, 94(5), 553–564. https://doi.org/10.1007/s11265-021-01724-3
8. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *ICASSP 2017 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2392–2396. https://doi.org/10.1109/ICASSP.2017.7952585
9. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017). CNN architectures for large-scale audio classification. *ICASSP 2017 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135. https://doi.org/10.1109/ICASSP.2017.7952132
10. Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), 7–19. https://doi.org/10.1109/TASLP.2014.2364452