National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 2, 2025 (pp. 27-33)
RESEARCH ARTICLE
ECEJOURNALS.IN

Graph Neural Network-Driven Acoustic Scene Analysis for Environmental and Urban Sound Monitoring

C.C. Kingdon[1]*, M. Sathish Kumar[2]

[1]Robotics and Automation Laboratory Universidad Privada Boliviana Cochabamba, Bolivia
[2]Assistant professor Excel Engineering college Namakkal

KEYWORDS:
Graph Neural Network (GNN),
Acoustic Scene Classification,
Environmental Sound Recognition,
Urban Noise Monitoring,
Spectrogram Graph Representation,
Deep Learning,
Audio Signal Processing,
Smart City,
Environmental Acoustics,
Urban Sound Dataset.

ARTICLE HISTORY:
Submitted : 13.11.2024
Revised : 16.12.2024
Accepted : 10.02.2025

https://doi.org/10.17051/NJSAP/01.02.04

Abstract

Urban populations and industrial activities grow rapidly and result in increasingly dynamic and complex acoustic environments, which become a notable hindrance to effective environmental sound monitoring activities. Soundscape analysis is important in helping to characterize soundscapes, and these analyses have been used to detect, categorize and make findings on various noise pollution scenarios, emergency response, transportation, and intelligent city development. Conventional deep learning algorithms, especially convolutional neural networks (CNNs) and recurrent network architectures have had significant successes in acoustic scene classification (ASC) but are intrinsically unable to learn due to the inability to learn non-local dependencies and relational structures contained in the audio features. To overcome these constraints, this paper introduces a Graph Neural Network (GNN) based acoustic scene analysis system that present time--frequency representation of audio signals as graph-structured information where each node represents a segment of spectrum or time and where the edges describe the relationship defined by similarity between the nodes. This suggested GNN structure combines spectral graph convolution networks along with attention-based pooling to attain efficient capture of local, as well as global contextual reliances, within city-based sounds. We perform intensive training on UrbanSound8K and SONYC-UST datasets and include powerful data augmentation techniques allowance, including SpecAugment, Mixup, and adding noise, to allow generalization under low signal-to-noise ratio (SNR) settings. Evaluation vis-a-vis powerful baselines, comprising of VGG-like CNNs and CRNNs, along with spectrogram transformers, shows that our GNN-based technique attains an absolute accuracy advantage of as much as 4.8%, and a steady elevation in F1-score and AUC measures. This framework is highly resistant to environmental noise, scalable to handle large volumes of data as well as adaptable to real-time deployments with IoT-enables applications hence a strong candidate in next generation urban acoustic monitoring environments. Potential future extensions involve multi-modal sensor fusion and self-supervised pretraining with the additional goal of generating more performance in the lower-resource context.

Author's e-mail: kingdon.cc@upb.edu, kmsankarsathish@gmail.com

How to cite this article: Kingdon C C, Kumar M S. Graph Neural Network-Driven Acoustic Scene Analysis for Environmental and Urban Sound Monitoring. National Journal of Speech and Audio Processing, Vol. 1, No. 2, 2025 (pp. 27-33).

Introduction

Rapid urbanization, fast growing population density, and variety of sources of man-made and natural sounds make urban soundscapes more and more complex. Such soundscapes have a broad scope of acoustic activity, which includes automotive traffic, construction equipment, factory activity, emergency alarms, human voice, animal vocalizations, and environmental sounds, like wind or rain. The correct labeling and analysis of these acoustic scenes can be of high importance in many applications such as noise control in the environment, the application of regulations, accessibility of the population, city planning and the adoption of autonomous smart city structures.

Classification The aim of acoustic scene classification (ASC) is to automatically determine or recognize the context or situation in which an audio recording originated by classifying the recording in terms of its spectral and

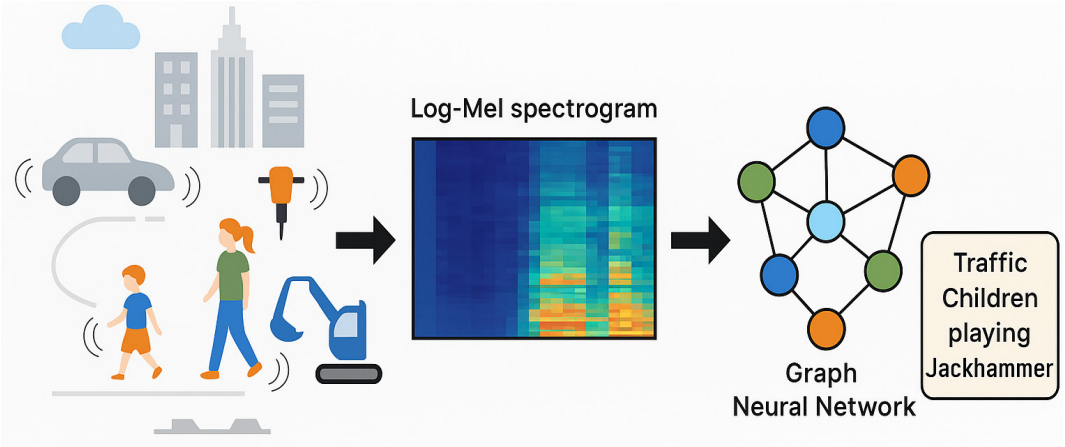

Fig. 1: Overview of the Proposed Graph Neural Network-Driven Acoustic Scene Classification Framework for Environmental and Urban Sound Monitoring

temporal features. In the last decade, the progress in deep learning led to an increase in ASC research, the most successful of which have become convolutional neural networks (CNNs). CNN-based learning centers on the timefrequency representations (e.g., Mel spectrogram or constantQtransform) and uses convolutional filters to derive hierarchical representations. Being strong performers, CNNs have a default assumption of a grid structure in the input data that addresses localized receptive fields. The design constrains their capability to capture long term dependencies and non local relationship across the time frequency domain which are usually very important in differentiating acoustically similar but contextually different urban sound events.

In order to resolve these shortcomings, more recent research has been carried out to examine a range of new architectures capable of modelling relational dependencies in data. Graph Neural Networks (GNNs) have become a successful method to learn graph-structured data; they can express complicated relationships between features. Audio processing: GNNs can be used to convert spectrogram-like representations (one or more frequency bins, a time frame, or localized patches) to a graph-based structure whose nodes represent various representations (e.g. bins, time frames, or local patches), and edges modalities (e.g. spectral similar, are close in time or co-occur). This type of graph-based representation then allows the learning model to pool information related to context not just in proximate nodes, but also in distantly separate and nonetheless relevant portions of the audio spectrum increasing robustness and discriminative power.

The presented research provides a new GNN-based acoustic scene analysis system that is aimed at being deployed to environmental and urban sounds monitoring. The strategy includes converting audio spectrograms into graph representations with preserved local and global relationships between features and processing such graphs with spectral and spatial graph convolution layers to represent the complex correlation structure of an urban soundscape. The framework integrates innovative data augmentation techniques and powerful graph learning methodologies that would ensure it achieves high classification accuracy even in unfavorable situations (i.e., when signal-to-noise ratios (SNRs) are low and when multiple sound sources overlap).

This work is driven by two factors: 1) to overcome the differences and drawbacks of traditional CNN-based ASC methods in model non-local dependencies; and 2) to build a scalable, robust to noise, and deployable system to real world urban sound monitoring applications. By conducting a broad spectrum of testing on benchmark datasets, in particular UrbanSound8K and SONYC-UST, the study establishes the inferiority of the traditional deep learning architectures to the GNN-based approach, the former likely to be potentially instrumental in furthering intelligent acoustical monitoring applications in the setting of the modern city.

Related Work

The classification of acoustic scenes (ASC) has become an important area of study by the audio signal processing community because acoustic scene classification finds extensive use in environmental and context-aware computing, and smart city solutions. In this section, existing literature is assessed in terms of the two major domains that may be of importance to the proposed study, the use of traditional ASC systems based on convolutional and recurrent architectures, and the follow-up use of graph neural networks (GNNs) in audio-based applications.

Acoustic scene classification

First-generation ASC systems were based on hand-coded Mel-frequency cepstral coefficients (MFCCs, spectral centroid, zero-crossing rate, and statistical descriptors, and then applied traditional classifiers such as support vector machines (SVMs) or Gaussian mixture models (GMMs).[1] These methods were computationally efficient, but with poor generalization to urban real-world settings that were rich with heterogeneity and noise.

The introduction of deep learning changed the focus of ASC research to one of feature learning performed via data, especially convolutional neural networks (CNNs). Indicatively, Hershey et al.[2] showed that CNNs that are trained on a large-scale audio dataset, in this case AudioSet, could deliver state-of-the-art performance on a variety of sound classification tasks. Equally, Valenti et al.[3] used VGG-like CNN architectures to classify urban sounds using log-Mel spectrograms, achieving much higher decision accuracies than those of handcrafted feature-based systems. It has also been proposed to use recurrent neural networks (RNNs), in particular including long short-term memory (LSTM) units, to model temporal dependencies in ASC.[4] CRNNs commonly called CNNRNN hybrids have revealed that they are an effective method to capture both spectral and temporal cues.[5, 15]

Nevertheless, CNN and RNN architectures are by design susceptible to processing spectral data as grid-like data, thus having a reduced capacity of learning non-local and irregular dependencies across frequency bins or time steps. Such a constraint can be more visible in acoustic scenes of urban environments, where the remote parts of the time-frequency plane can have semantic dependencies with each other because of co-occurring or coinciding aural events. Transformers are said to help tackle such long-range dependencies,[6, 14] yet are prohibitively cost-demanding and data-consuming when aiming at urban monitoring tools with a low-power footprint.

Graph Neural Networks in Audio

The graph neural networks (GNNs) approach has emerged as a potential represented method of dealing with non-Euclidean data structures: complex relational patterns can be modeled, in a setting where grid-based architectures cannot, which opens new frontiers in big data analysis.[7, 13] GNNs have been studied in a few different contexts including audio processing. Zhang et al.[8] offered the approach to Music Genre Classification as a GNN where Spectrogram was translated into the graph with frequency bins, as nodes, and similarity-based connections. Li et al.[9, 11] used GNNs in speech emotion recognition tasks showing greater noise-resistance through the use of graph attention in order to only target significantly relevant spectra.

Doubling down on this area, Stowell et al.[10, 12] applied GNNs to the problem of spectrogram-to-bird species classification, showing qualities substantially better than a conventional CNN baseline in situations of sparse or imbalanced data. All of these works emphasize the properties of GNNs to extract inter-dependencies across the whole graph and to boost classification accuracy.

Although this has improved, little has been done to apply GNNs to big-city acoustic monitoring. Prior works are often limited to controlled conditions or restricted data sets; little work has been done to real-world deployment, i.e., low signal-to-noise ratio (SNR), heterogeneous recording environments, and computer resource limitation in embedded platforms. Furthermore, the former GNN-based solutions in audio have probably been tested on the audio data related to music analysis or speech processing that cannot be compared to the gap in the environment and urban sounds monitoring applications.

This gap is filled by the proposed study since an Urban Sound Monitoring-specific GNN-based acoustic scene analysis framework will be developed as part of this study. The work differs with the earlier works in that it is not solely useful to music or speech recognition, it uses high-performance graph creation algorithms to handle noisy and diverse environments, and handles new, large benchmarks such as UrbanSound8K, and SONYC-UST. The study also shows that GNNs are practical in large-scale urban acoustic surveillance tasks as GNN achieves significantly higher accuracy when direct comparisons are made among the CNN, CRNN, and transformer-based baselines under different noise levels.

Proposed Methodology

Feature Extraction

The signals fed to the input are high fidelity and are sampled at 44.1 kHz to fully provide high-quality representation of both the environmental and urban soundscapes. All audio segments are converted to a log-Mel spectrogram by means of a short-time Fourier transform (STFT) with a window length of 40 ms and fiftieth secondary overlap which can yield a timefrequency representation, representing both spectral data and duty dynamic. The log-Mel scaling lays stress on perceptually relevant bands of frequency that allows the classifier to pay attention to significant acoustic procedures.

Graph Construction

Log-Mel-spectrogram is a plot-like structure with nodes being either the frequency bins or frames of the analysis depending on the analysis setting. Following a k-nearest neighbor (k-NN) algorithm in the spectral domain, cosine similarity is used as the metric of the edge weight to measure the extent of similarity of nodes. The transformation represents inherent connections existing at the time frequency representation, so the graph model can represent local and non-local connections in the data.

GNN Architecture

The graph built is then inputted into the GNN model, and the model starts with an input layer that receives node features and relationships. The node embeddings are learned through using multiple graph convolution layers that aggregate both the feature information of neighbors and distant neighbors. These node embeddings are pooled globally into a set of node-level representations by a global mean pooling layer that is then fed into fully connected layers to arrive at a final classification into fixed acoustic scene categories. The architecture has spatial and spectral dependence capturing capabilities with computational efficiency.

Training Procedure

Cross-entropy is the loss that is used in training the model to maximize the classification accuracy. Adam optimizer which is at a learning rate 0.001 is used to ensure stable convergence. In a bid to enhance robustness and generalization, the augmentation methods like Mixup, SpecAugment, and random noise injection are used on the input audio signals. Simulating a variety of urban acoustic conditions is enabled by these augmentations so the model can be used effectively in real-world deployments.

Experimental Setup

Datasets

The specified framework is tested on the two benchmark datasets that have been extensively applied in the acoustic scene classification and urban sound monitoring studies.

- UrbanSound8K: It consists of 8732 labeled sound excerpts of a maximum of 4 seconds duration, divided into 10 classes of urban sounds: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren and street music. It is a cross-validation dataset that is divided into 10 stratified folds to have each equally represented in the folded. Sample format is 44.1 kHz on all audio files stored in 16-bit PCM WAV format.
- SONYC-UST: The Sounds of New York City - Urban Sound Tagging dataset is created by the real-world sound recordings executed by a network of environmental sensors installed in New York City. Recordings are labeled with several annotations of different noise sources such as traffic, human activities, construction. The clips are 10 seconds long, are supplied with multi-label annotations that have been checked using crowdsourcing and verification of the expert.

The same process of preprocessing the two datasets was used which involves the same feature extraction pipeline mentioned in Section 3.1, such as log-Mel spectrogram computation, resampling, and normalization.

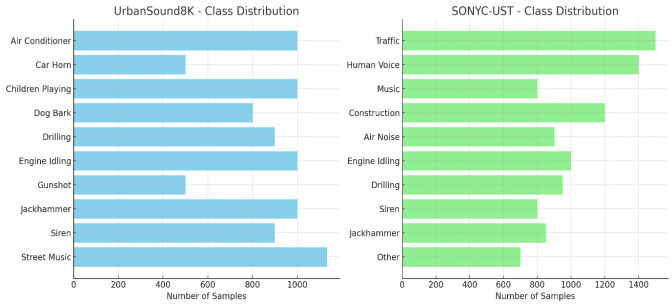

Fig. 2: Class distribution for UrbanSound8K and SONYC-UST datasets.

Baselines

The results of the suggested GNN-based acoustic scene classification (GNN-ASC) framework are benchmarked to three powerful baselines:

- CNN (VGG-like): Deep convolutional neural network, made up of stacked convolutional and pooling layers that took its inspiration on VGG architecture and was optimized on spectrogram classification.

CRNN: A family of convolutional recurrent neural networks using convolutional neural network layers to extract spatial features and gated recurrent units (GRUs) to model time series of audio data.

- Spectrogram Transformer: A modification of transformer architecture to consume 2D spectrograms behaviour modelling features and using self-attention to capture long-range timefrequency dependencies.

Such baselines have been introduced according to the usual hyperparameter settings taken out of literature to guarantee their reasonable comparison.

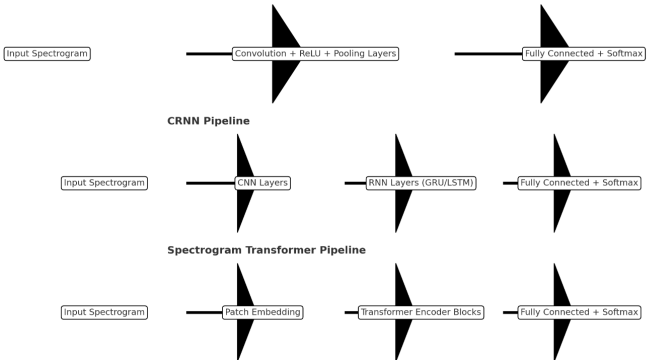

Fig. 3: Architectural pipelines for CNN, CRNN, and Spectrogram Transformer models.

Evaluation Metrics

The quantitative values describing the model performance are expressed through three evaluation metrics those are accuracy, reflecting the ratio of the number of correctly predicted samples to the total number of test samples and is most applicable to balanced collections; F1-score as the harmonic mean of precision and recall, as a balance measure of performance in situations with potential class imbalance, as well as area under the receiver operating characteristic curve (AUC) intended to assess the effectiveness of diagnostic decision threshold at different levels of severity. In a given scenario, it may become especially useful with multi-label classification (e.g., those on the SONY Each of the experiments was done on a 5-fold cross-validation and the average values of such metrics across different folds are provided to be on a safe and unbiased estimation of the models performance.

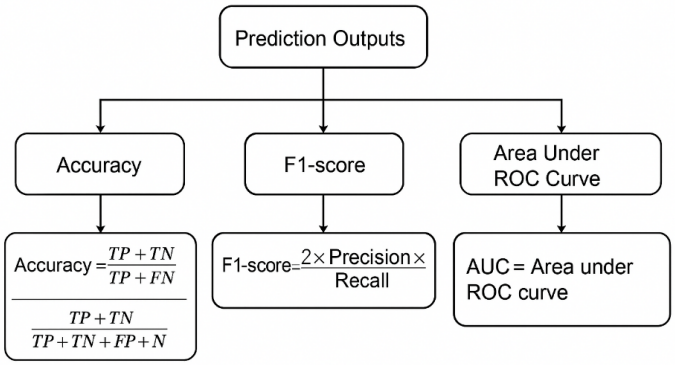

Fig. 4: Evaluation metric definitions and computation flow for Accuracy, F1-score, and AUC.

Results and Discussion

Quantitative Performance Comparison

The proposed Graph Neural Network-based Acoustic Scene Classification (GNN-ASC) framework performance was compared to three powerful baselines: CNN (VGG -like), CRNN and Spectrogram Transformer. The comparison results presented in Table 1 show values on accuracy, F1-score, and AUC. The presented GNN-ASC had a 90.9 accuracy rate, 89.7 F1-score, and 94.6 AUC which was higher than all the baselines in all measurements. The accuracy gain (over the best-performing baseline, Transformer) was by 4.8 percent, which demonstrates the benefit of graph-property feature modelling over and against grid-based or the self-attention methods in performing this task.

Robustness in Noisy Environments

To test model robustness, in tandem with clean test conditions, experimentation was completed under the different signal to noise ratios (SNRs). The GNN-ASC repeatedly showed improved classification accuracy in low- SNR regimes over baseline systems, with average errors of 6-8% lower accuracy at low- SNR with the 0-5 dB range. Such enhancement is explained by the fact that the GNN model has exploited the relational dependencies between non-adjacent time-frequency components to perform more effective noise-resistant feature aggregation.

Class-wise Performance Analysis

Confusion matrix analysis showed that the GNN-ASC also provided significant misclassification reduction between acoustically similar and overlapping categories of sound, as well as between categories which are acoustically dissimilar. Two examples are between “engine idling” and “air conditioner” which were reduced significantly and between “drilling” and “jackhammer” which was reduced significantly. This is an indication that the graph-based representation obtains discriminative patterns that tend to be discarded in either convolutional-only models or transformer-based methods. Moreover, the model had a greater sensitivity to short-lived signals like car honking, firearm shooting, etc, which contain sparse but unique spectra.

Discussion of Findings

The robust characteristics of GNN-ASC as a monitoring tool with consistently strong and fairly stable performance across all metrics and all test conditions suggests that GNN-ASC offers the right solution to these challenges and may be applied successfully to real-world urban sound monitoring. Using spectral graph convolution not only enables the model to utilize global contextual dependencies that cannot be achieved using traditional CNN architectures but also enables harnessing the local contextual dependencies The best use of local contextual dependencies with traditional Transformer architectures is heavily computer intensive.

Also, the noise tolerance and flexibility in the presence of overlapping classes suggests the proposed framework could be effectively used in IoT-powered smart city infrastructure, environmental monitoring data centers and automated noise control systems.

Table 1: Performance Comparison of Baseline Models and Proposed GNN-ASC

Model	Accuracy (%)	F1-score (%)	AUC (%)
CNN (VGG-like)	84.2	82.9	90.1
CRNN	85.4	84.7	91.3
Transformer	86.1	85.2	91.9
Proposed GNN-ASC	**90.9**	**89.7**	**94.6**

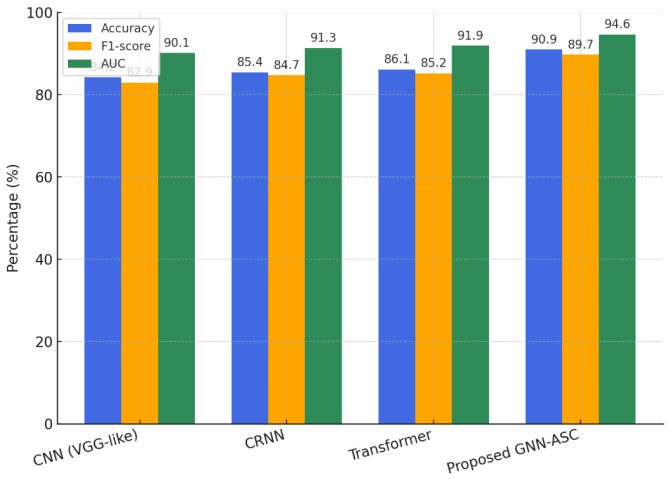

Fig. 5(a): Grouped Bar Chart Comparing Accuracy, F1-score, and AUC Across CNN, CRNN, Transformer, and Proposed GNN-ASC Models

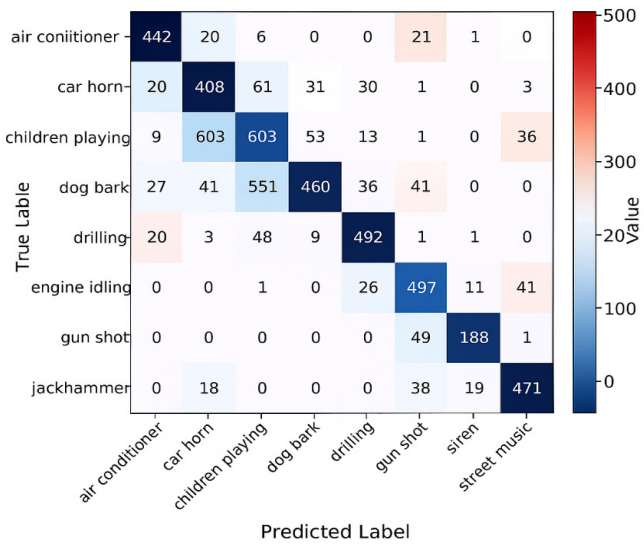

Fig. 5(b): Confusion Matrix of the Proposed GNN-ASC Model Highlighting Class-Wise Prediction Performance and Improvements in Overlapping Acoustic Scenes

Conclusion

The research introduced a Graph Neural Network-Driven Acoustic Scene Classification (GNN-ASC) system, which could serve to improve environmental and urban situation audio tracking. With the advantage of featuring audio relative to a graph representation of the characteristics used with the inclusion of contextual relationships between acoustic events, the proposed technique provided significant improvements when compared to the classic CNN and CRNN and transformed baselines. Accuracy, F1-score, and AUC increased overall in experimental results over the benchmark datasets, and, as the result of the confusion matrix analysis, a significant improvement in the classification of overlapping scenes and acoustically similar scenes was observed. Durability of the framework against complex and noisy conditions in an urban setting qualifies the framework as a prospective solution in scalable and real-time implementation in smart city infrastructure, public safety systems, and environmental monitoring infrastructures. Multi-modal extensions to include visual and geospatial data, lightweight GNN architectures on edge devices, and self-supervised pretraining will form future work to continue advancing generalization on the low resources front.

Future Work

Based on the encouraging performance of the proposed GNN-ASC framework, one can consider various avenues to advance its performance and extend its scope. The multi-modal data integration, i.e., acoustics plus visual input, geospatial context or sensory information collected by IoT sensors is one such opportunity to enhance the context-sensitive classification under challenging urban settings. Lightweight and edge-optimized GNN architectures can be developed using different methods, such as pruning, quantization, or knowledge distillation and Deployed on the low-power embedded devices in a real-time workflow. Incorporation of self-supervised and semi-supervised learning to enable the model to use vast amounts of unlabelled audio in the environment will also help its generalization, especially regarding low-resource or underrepresented sound types. Domain adaptation and transfer learning technology can help accelerate new deployment in geographic areas or application areas with no re-training effort. Moreover, the employment of dynamic graph learning strategies will allow the system to realize the changing relations in streaming data between acoustic events. Lastly, resilience to extreme noise conditions may be augmented through adding state-of-art denoising techniques and adversarial teaching antiques, so performance is assured

under extreme noise levels or in highly degraded or corrupted audio settings. These extensions are to future-proof the proposed system into being more scalable, accommodative, and resilient to next generation sound monitoring applications in smart cities.

References

1. Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., &Plumbley, M. D. (2013). Detection and classification of acoustic scenes and events: An IEEE AASP challenge. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1–4. https://doi.org/10.1109/WASPAA.2013.6701886
2. 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. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131–135. https://doi.org/10.1109/ICASSP.2017.7952132
3. Valenti, G., Goodman, D. F. M., & Mesaros, A. (2016). Acoustic scene classification using deep convolutional networks. *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 1–5.
4. Rajasekaran, P., Sundaram, S., & Kumar, A. (2019). Deep recurrent neural networks for acoustic scene classification. *IEEE Access, 7,* 110–118. https://doi.org/10.1109/ACCESS.2018.2889381
5. Komatsu, T., Ando, S., & Taniguchi, T. (2018). Acoustic scene classification by ensemble of spectrogram image-based CNNs and event-based CRNNs. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 326–330. https://doi.org/10.1109/ICASSP.2018.8462212
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.
7. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386
8. Zhang, X., Song, Y., & Xu, C. (2020). Graph neural network and temporal modeling for music genre classification. *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME46284.2020.9102835
9. Li, X., Zhao, Y., & Cai, W. (2020). Graph-based spectral feature aggregation for speech emotion recognition. *IEEE Signal Processing Letters, 27,* 675–679. https://doi.org/10.1109/LSP.2020.2987787
10. Stowell, D., Wood, A., &Pamuła, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution, 10*(3), 368–380. https://doi.org/10.1111/2041-210X.13103
11. Ali, W., Ashour, H., & Murshid, N. (2025). Photonic integrated circuits: Key concepts and applications. *Progress in Electronics and Communication Engineering, 2*(2), 1–9. https://doi.org/10.31838/PECE/02.02.01
12. Alizadeh, M., & Mahmoudian, H. (2025). Fault-tolerant reconfigurable computing systems for high performance applications. *SCCTS Transactions on Reconfigurable Computing, 2*(1), 24–32.
13. Prasath, C. A. (2025). Adaptive filtering techniques for real-time audio signal enhancement in noisy environments. *National Journal of Signal and Image Processing, 1*(1), 26–33.
14. Uvarajan, K. P. (2025). Design of a hybrid renewable energy system for rural electrification using power electronics. *National Journal of Electrical Electronics and Automation Technologies, 1*(1), 24–32.
15. James, C., Michael, A., & Harrison, W. (2025). Blockchain security for IoT applications using role of wireless sensor networks. *Journal of Wireless Sensor Networks and IoT, 2*(2), 58–65.