Anusandhanvallari
Vol 2024, No.1
September 2024
ISSN 2229-3388

Attention-Enhanced Deep Learning for Urban Environmental Sound Classification

Syed Sibtain Khalid[1*], Safdar Tanweer[2], Farheen Siddiqui[3], Mohd Abdul Ahad[4], Rahbre Islam

[1*]Department of computer Science and Engineering, Jamia Hamdard, New Delhi, India

[2]Department of computer Science and Engineering, Jamia Hamdard, New Delhi, India

[3]Department of computer Science and Engineering, Jamia Hamdard, New Delhi, India

[4]Department of computer Science and Engineering, Jamia Hamdard, New Delhi, India

E-mail: sibtain1977@gmail.com

Abstract

Acoustic environmental noise classification is an important task for reliable and intelligent acoustic systems. Furthermore, it is still challenging due to non-stationary nature, overlapping acoustic pattern, background interference and availability of limited tagged dataset of environmental noise. This paper presents a robust approach for urban environmental noise classification with the help of deep learning techniques using UrbanSound8K dataset. This proposed approach uses logarithmic-Mel spectrograms with data augmentation and Convolutional Neural Network that is enhanced by attention mechanisms which captures discriminative time-frequency features. To enhance computational efficiency and reproducibility, features were pre-computed and cached without altering the learning process. Official 10-fold cross-validation protocol is used to stop data leakage and ensure comparison with previous studies. Experimental results show an average classification efficiency of about 91.14%, outperforming traditional CNN baseline models while the light weight architecture is useful in many practical deployments. Detailed analysis of each class further unfolds the good recognition accuracy of impulsive and harmonic acoustic sounds, furthermore highlighting the challenges in recognition of similar categories of acoustic signal. The result indicates that CNN combined with efficient preprocessing gives a practical and scalable real time urban sound monitoring model. In future, use of large scale dataset can further enhance the performance.

Keywords: Environmental sound classification, Deep learning, Convolutional neural networks, Log-mel spectrogram, intelligent audio analysis

1. Introduction

An increasing urbanization causing different kinds of environmental noises affecting many aspects of life, such as health, safety etc. Hence intelligent monitoring of acoustic environmental noise is becoming an important task for implementing smart city concepts. Urban environmental noise is highly unstructured and has substantial variability in its characteristics, which leads to making it a challenging task to detect and identify different kinds of environmental noise available in the surroundings. Conventional approaches generally use Mel-frequency cepstral coefficients (MFCC), zero crossing rate, spectral characteristics, Short Time Fourier Transform (STFT) followed by conventional classifier. These techniques result in moderate classification accuracy due to limited generalizing capabilities. The emergence of deep learning [16] enables data driven feature learning directly from the time frequency representation of acoustic signals. Convolutional Neural Network (CNN) shows good accuracy by capturing local spectral and temporal patterns from spectrograms [1,17].

Despite advancement still some challenges occur. As deep learning models treat all spectral regions equally, which creates problems in identifying the noise class because only certain time-frequency components are relevant. Another problem is the deployment of intelligent systems that also requires computational efficiency and

reproducibility during training besides high accuracy.

This paper proposes an attention enhanced deep learning framework for urban environmental noise classification to combat the above said problem. In this approach we have used log-mel spectrograms to represent acoustic signals in a perceptually meaningful time-frequency demonstration. A convolution neural network with channel wise-attention is designed to automatically capture useful spectral features while suppressing other background components. This attention enhancing feature enables the design to better distinguish similar kinds of environmental acoustic noises. We have pre-computed the feature and cached prior to training which ensures deterministic preprocessing, reducing redundant computation that are beneficial for scalable intelligent system deployment. The block diagram of the proposed model is shown in *Fig.1*.

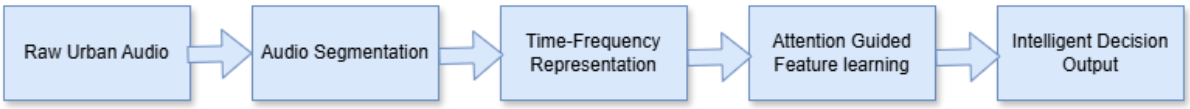

Fig.1 Overview of the proposed urban environmental sound classification.

We have utilized UrbanSound8K dataset with official 10-fold cross validation protocol to maintain leakage free performance comparison [2]. Experimental results show classification accuracy of 91.14%, which is robust across diverse acoustic environmental noise. We also analyzed class wise behavior to further investigate the strength and challenges in recognition of acoustic urban environmental noise.

2. Related Work

2.1 Environmental Sound Classification

Environmental sound classification (ESC) is an audio recognition subfield that identifies and categories different types of environmental acoustic signals [3]. In the modern era of artificial intelligence and robotics it has got potential attention due to high end utility for surveillance, smart city development, and context-aware systems. Environmental sounds are generally unstructured, non-stationary, and often occur multiple acoustics simultaneously, making automatic recognition a challenging task [4,19]. Infancy ESC systems are based on handcrafted acoustic features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid, zero-crossing rate, and temporal energy. These features were utilized along with the traditional classifiers such as support vector machines (SVMs), Gaussian mixture models (GMMs), and k-nearest neighbors (k-NN) [5]. Due to diverse and complex urban acoustic patterns, the performance of the traditional technique was limited.

The ESC performance significantly improved with the emergence of deep learning that enables automatic feature learning from time–frequency representations. The use of Convolutional neural networks (CNNs) to the spectrogram-like inputs results in a powerful approach for recognizing different acoustic events because they effectively model local spectral and temporal correlations. Different research reveals various kinds of spectrogram representations such as log-mel spectrograms, constant-Q transform (CQT), and gamma tone filter banks, with log-mel features that provide a good balance between perceptual relevance and computational efficiency. UrbanSound8K dataset is providing a benchmark for standardizing evaluation protocols and enabling fair comparison among different types of deep learning models [6].

The CNN-based approaches still have challenges in distinguishing similar acoustic classes and handling intense interference by the urban acoustic surroundings that are available most of the time.

2.2 Attention Mechanisms in Audio Deep Learning

Recently attention mechanisms are utilized in deep learning to enhance model performance by considering the most informative parts of the input. This mechanism is originally developed for natural language processing and also successfully implemented for computer vision and audio analysis tasks [18]. The use of attention modules in environmental sound classification are to reweight feature maps along temporal, spectral, or channel dimensions, that results in enhanced discriminative patterns while suppressing irrelevant components of the acoustics [7].

The use of channel attention mechanisms, such as squeeze-and-excitation (SE) blocks, effectively models the

interdependencies of feature channels with minimal computational overhead. For longer acoustic segments temporal attention has also been explored. These mechanisms enhance robustness and interpretation of acoustic events with multiple sources. However, use of attention in ESC is limited and without considering computational effectiveness as well as practical deployment constraints [8].

2.3 Efficient and Reproducible ESC Frameworks

The need of real-world intelligent systems is not only classification efficiency but also computational efficiency and reproducibility. Training of deep audio models can be more resource-intensive, in case feature extraction is repeatedly computed during each training epoch. In the recent past it suggests the benefit of pre-computing time-frequency features that decreases redundant processing and also ensures consistent input available across training runs. These strategies are very useful in large-scale experiments and cross-validation protocols, yet they are considered as implementation details rather than integrated into system design discussions [9].

Additionally, the real-time acoustic event monitoring system can be lightweight but accurate ESC models. The optimization of model complexity, performance, and computational cost is an open research problem for intelligent acoustic analysis.

2.4 Research Gap and Motivation

The literature review reveals that CNN-based models using log-mel spectrograms structures a strong foundation for urban environmental sound classification [10]. The use of attention mechanisms improves feature discrimination, but performance and efficiency of the integrated ESC system is not fully analyzed. Furthermore, less preference has been given to reproducible preprocessing pipelines that reduce computational redundancy while maintaining methodological rigor.

To address these gaps, we have proposed an attention-enhanced CNN framework combined with preprocessing strategy for an efficient and deterministic spectrogram. The technique aims to enhance classification robustness in recognizing urban acoustic signals while considering computational practicality for intelligent monitoring applications.

3. Proposed Methodology

In this section we will discuss proposed attention-enhanced deep learning techniques for urban environmental sound detection. The proposed system consists of three stages: feature extraction, attention-guided convolutional feature learning and prediction stage. This section describes the proposed attention-enhanced deep learning framework for urban environmental sound classification.

3.1 Acoustic Feature Extraction

Time-frequency representation is used to capture temporal and spectral patterns. A discrete time acoustic signal x[n] is converted into time-frequency representation using Short Time Fourier Transform (STFT) as shown in *equation (i).*

$$X(t,f) = \sum_{n=-\infty}^{\infty} x[n]w[n-t]e^{-2\pi fn} \quad \text{……………………………………………………………}(i)$$

Where w[n] is the window function, f and t represent frequency and time index respectively. The magnitude spectrum is now travel through the mel filter bank to get mel-band energies that is evaluated using *equation (ii).*

$$M_k(t) = \sum_{f=f_{min}}^{f_{max}} |X(t,f|^2 H_k(f) \quad \text{……………………………………………………………}(ii)$$

Where $H_k(f)$ is the k-th triangular mel filter. Log-mel spectrogram is obtained by applying logarithmic compression as indicated by *equation (iii).*

$$S_{log}(k,t) = log(M_k(t)+\epsilon) \quad \text{………………………………………………………………………}(iii)$$

Where ϵ is a constant to minimize numerical instability. The obtained log-mel spectrogram $S_{log}(k,t)$ is utilized to the neural network as input. These spectrogram features are pre-computed and cached prior to model training to ensure efficiently and reproducibility.

3.2 Attention-Enhanced Convolutional Neural Network

A convolutional neural network (CNN) augmented with channel-wise attention is used in this model to improve discriminative feature learning. A series of convolutional, batch normalization, activation, and pooling layers are utilized by the CNN to extract temporal-spectral features from the log-mel spectrogram [11].

First, the global average pooling is considered to aggregate spatial information for each channel as given by:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} F_c(i,j) \quad \text{.....(iv)}$$

The output of channel descriptor z is passed through a lightweight gating network that are made up of two fully connected layers and nonlinear activation functions to generate attention weights:

$$\alpha=\sigma(W_2\delta(W_{1z}) \quad \text{.....(v)}$$

Where W_1 and W_2 are weight matrices for learnings, $\delta(\cdot)$ is a ReLU activation function, and $\sigma(\cdot)$ is the sigmoidal activation function. The modified feature maps are now obtained by channel-wise scaling as indicated in equation *(v)* [12].

$$\hat{F}_c = \alpha.Fc \quad \text{.....(vi)}$$

This technique enables the network to consider informative acoustic patterns while suppressing less relevant components.

3.3 Classification Layer

The attention-enhanced features produce a more compact feature vector by applying additional convolutional layers followed by global average pooling on it. A fully connected layer with softmax activation generates posterior probabilities over N sound classes as indicated by equation *(vii)*:

$$\hat{y}_i = \frac{e^{u_i}}{\sum_{j=1}^{N} e^{u_j}} \quad \text{.....(vii)}$$

Where u_i is the output logit corresponding to class *i*.

3.4 Training Strategy

This model is trained by the cross-entropy loss function as described by the equation *(viii)*.

$$CE = -\sum_{i=1}^{N}(y_i * \log(\hat{y}_i)) \quad \text{.....(viii)}$$

Where y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability. Data augmentation techniques like time shifting and additive noise are imposed to improve generalization. Training is performed using the official 10-fold cross-validation protocol of UrbanSound8K to ensure fair and leakage-free evaluation.

4. Experimental Setup

4.1 Dataset

Here we have used the UrbanSound8K dataset that contains 8732 labeled acoustic environmental noise up to 4 seconds duration of 10 different urban sound classes. It includes air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. The audio recordings show high variability in background noise, recording conditions, and sound intensity, making the dataset useful for analyzing and preparing robust classification models. Fig.2 shows an example of an acoustic sample taken from the dataset and represented in the time domain.

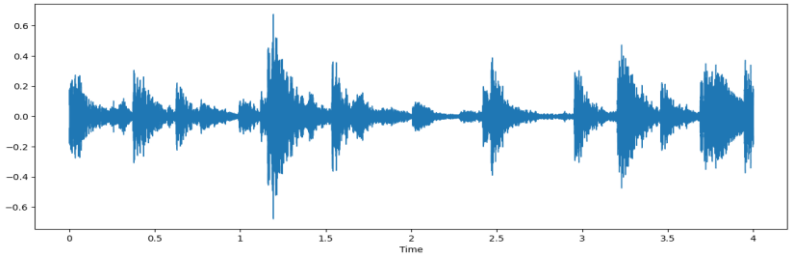

Fig.2: Time domain representation of an acoustic signal sample

UrbanSound8K has a predefined **10-fold cross-validation** split, where acoustic signals generated by the same source are grouped within the same fold to prevent data leakage [13]. It ensures fair and standardized comparison with the existing models. Fig.3 shows the distribution of different classes of acoustic signals present in the dataset.

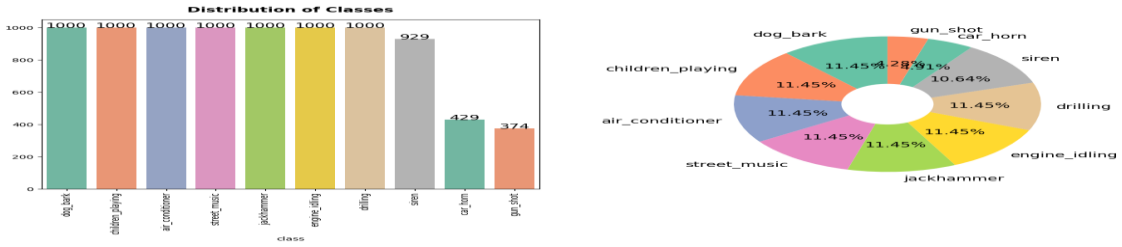

Fig.3: Distribution of different classes of UrbanSound8K dataset

4.2 Feature Extraction Parameters

Each acoustic signal is sampled using a uniform sampling rate of 22.05 kHz and truncated to a fixed duration of 4 seconds. Log-mel spectrograms are evaluated with STFT window size of 1024 samples, hop length of 512 samples, number of mel filter banks of 128, frequency range of 0-11 kHz and logarithmic compression with numerical stability constant $\epsilon=10^{-6}$.

4.3 Network Architecture

The model is made up of multiple convolutional blocks followed by channel-wise attention modules. The convolutional block performs convolution, batch normalization, ReLU activation, and max pooling. Channel attention is achieved by squeeze-and-excitation mechanism to adjust recalibrate feature map importance.

Global average pooling is utilized for the final fully connected classification layer. The model is designed to optimize a balance between classification accuracy and computational efficiency that makes it suitable for intelligent real-time monitoring systems [14].

4.4 Training Configuration

Adam optimizer is used to train the model with an initial learning rate 0.001.The model is trained using with an initial learning rate of **0.001**. Step decay scheduler reduces the learning rate when validation performance plateaus. A batch size of **32** is used during training.

Each fold is trained for 30 training epochs to minimize cross-entropy loss function. Over fitting of data is prevented by early stopping based on validation loss [15]. Data augmentation is achieved and employed during training with the help of random time shifting and additive Gaussian noise that improves generalization in varying acoustic conditions. The final performance is evaluated as the average accuracy across all folds.

4.5 Evaluation Metrics

Overall accuracy is measured as the ratio of correctly predicted class to the total number of samples. Confusion matrix is used to analyze the model behavior across the different classes of acoustic noises.

5. Results and Discussion

5.1 Overall Classification Performance

The proposed system achieves a mean classification accuracy of 91.14%. The attention-enhanced deep learning technique is used on the official 10-fold cross-validation protocol of the UrbanSound8K dataset. The result demonstrates a strong generalization across different kinds of urban acoustic signals. The integration of channel-wise attention with convolutional features leads to the model which is capturing better discriminative spectral-temporal patterns compared to standard CNN architectures as indicated by the results. The particularly notable improvement is to recognize background acoustical interference within the dataset.

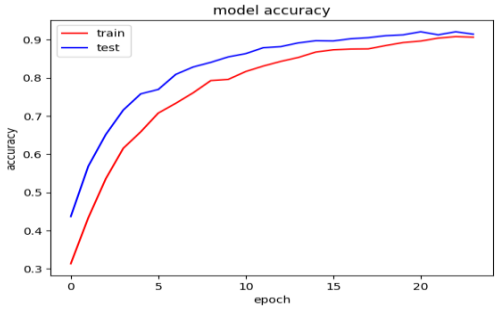

Fig.4: Accuracy curve of the model

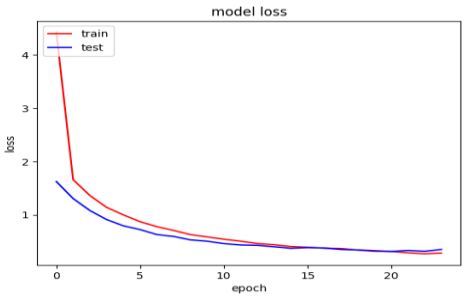

Fig.5: Loss curve of the model

Fig.4 shows the accuracy obtained during training and validation in 30 epochs while Fig.5 shows loss curve for the proposed model.

5.2 Comparison with Baseline Approaches

The performance of the proposed model is compared with the conventional CNN-based approaches that use a log-mel spectrogram without attention mechanism. The CNN baseline model achieves an accuracy range of 85-89% on UrbanSound8K dataset under the same cross-validation protocol. Hence the proposed model is surpassing these baseline models that indicates attention enhanced feature recalibration enhances the model performance by surpassing irrelevant components. It also maintains a relatively lightweight design. It is suitable for real-time intelligent monitoring systems that require accuracy as well as efficiency.

5.3 Class-wise Performance Analysis

Different classes of acoustics were analyzed on the basis of a confusion matrix that demonstrates a detailed class-wise analysis was conducted using the confusion matrix. Fig.6 shows the normalized confusion matrix obtained for the proposed attention enhanced model. The highest accuracy is obtained for children playing acoustic signal whereas least accuracy is for gun shot and jack hammer.

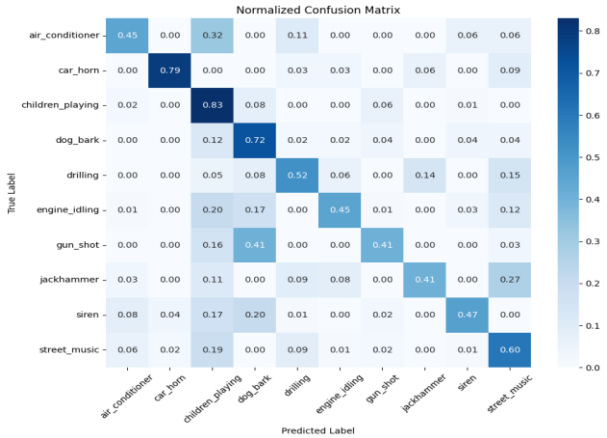

Fig.6: Normalized confusion matrix of classifier

The model demonstrates strong recognition performance for **high spectral contrast**, **distinct temporal structure**, **wide frequency distribution**, **strong energy concentration** and clear discriminative patterns that align well with convolutional receptive fields such as dog bark, street music, children playing. These acoustic sounds have distinctive spectral structures that are more effectively captured with the help of attention-enhanced mechanisms [16]. Whereas gunshot and car jack hammer shows a lower recognition efficiency. Gunshots have very short duration signals which occupy a small portion of the spectrogram in time, hence its temporal saliency reduces within the given fixed time frame. Jack hammers contain narrow band harmonic energy that spectrally overlap with other such events and make it difficult to detect. These factors make them comparatively difficult to classify. Fig.7 shows the ROC curve for each class of acoustic signal of UrbanSound8K for the proposed model.

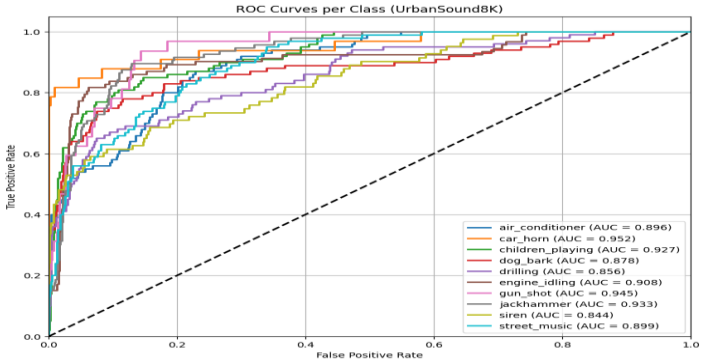

Fig.7:ROC curve for different classes

Table-I represented here shows a detailed view of different classes of acoustics precision, recall,f1-score and support obtained during the experiment.

Table-I

Classification Report:

	precision	recall	f1-score	support
air_conditioner	0.9151	0.9557	0.9349	203
car_horn	0.9750	0.9070	0.9398	86
children_playing	0.7953	0.9344	0.8593	183
dog_bark	0.8838	0.8706	0.8772	201
drilling	0.9087	0.9175	0.9130	206
engine_idling	0.9543	0.9741	0.9641	193
gun_shot	0.9333	0.7778	0.8485	72
jackhammer	0.9749	0.9327	0.9533	208
siren	0.9408	0.9636	0.9521	165
street_music	0.9234	0.8391	0.8793	230
accuracy			0.9141	1747
macro avg	0.9205	0.9073	0.9122	1747
weighted avg	0.9169	0.9141	0.9142	1747

5.4 Effect of Attention Mechanism

The integration of channel-wise attention with the baseline CNN model significantly improves classification robustness. The network prioritizes informative spectral cues by adaptively reweighing feature maps and reducing the influence of less relevant components. This technique enhances feature discrimination without adding much complexity in the system.

The performance analysis of the proposed system indicates that attention modules result in an effective trade-off between accuracy and computational cost, making them more suitable for intelligent audio analysis systems that can perform well under practical resource constraints.

5.5 Computational Considerations

The log-mel spectrograms were pre-computed and cached before model training to enhance training efficiency and reproducibility. In this way we have eliminated redundant feature extraction during each epoch and also significantly decreased computational overhead. At the same time this optimization technique also ensures the learning process or evaluation protocol does not alter, that makes the performance comparisons fair and methodologically sound. The outcome of the result suggests a balanced **accuracy**, **efficiency**, and **practical deployability** that are required for expert systems used in real-world smart city scenarios.

5.6 Discussion

The results show that attention-guided CNN models combined with acoustic features give a concrete solution for urban environmental sound classification. The system performance achieves a high level of accuracy but still challenges remain in distinguishing acoustically similar sound. In future we can improve the model by integrating transformer based temporal modelling or transfer learning on larger audio datasets. Finally the proposed approach represents a practical and scalable step in getting an intelligent environmental acoustic classification system.

6. Conclusion

The paper presented an idea of attention-enhanced deep learning technique for urban environmental sound classification. It supports the concept of intelligent acoustic monitoring systems for various practical purposes. In this approach log-mel spectrograms are used to represent time-frequency features in a more meaningful manner and incorporate it with channel-wise attention mechanisms to improve discriminative feature learning. It suppresses irrelevant acoustics in background and emphasizes on informative spectral components to enhance the model robustness in a complex urban acoustic environment.

The UrbanSound8K benchmark dataset and the official 10-fold cross-validation protocol is used to analyze fair and standardized performance assessment. Experimental results show an accuracy of 91.14% of attention-guided feature recalibration, which is fairly better in comparison to conventional CNN baseline model. Furthermore, the use of pre-computed spectrogram features enhances efficiency and reproducibility that are required for intelligent scalable systems.

The analysis of each class reveals strong performance for distinctive urban acoustic events and highlighting the ongoing challenges in classifying the narrow band harmonic energy acoustics that spectrally overlap with other such events. These findings suggest although attention mechanisms enhance feature discrimination yet improvement in modeling long-range temporal dependencies can help to address residual confusion.

References

[1] Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE access*, *11*, 106620-106649.

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

[3] Bansal, A., & Garg, N. K. (2022). Environmental Sound Classification: A descriptive review of the

literature. *Intelligent systems with applications*, *16*, 200115.

[4] Tanweer, Safdar, Abdul Mobin, and Afshar Alam. "Environmental noise classification using LDA, QDA and ANN methods." *Indian Journal of Science and Technology* 9.33 (2016): 1-8.

[5] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, *10*, 122136-122158.

[6] Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).

[7] Mu, Wenjie, et al. "Environmental sound classification using temporal-frequency attention based convolutional neural network." *Scientific Reports* 11.1 (2021): 21552.

[8] Yang, Chao, et al. "ResNet based on multi-feature attention mechanism for sound classification in noisy environments." *Sustainability* 15.14 (2023): 10762.

[9] Schluter, J., & Gutenbrunner, G. (2022, August). Efficientleaf: A faster learnable audio frontend of questionable use. In *2022 30th European signal processing conference (EUSIPCO)* (pp. 205-208). IEEE.

[10] 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.

[11] Dong, Shaojiang, et al. "Environmental sound classification based on improved compact bilinear attention network." *Digital Signal Processing* 141 (2023): 104170.

[12] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[13] Guo, Jinming, et al. "A deep attention model for environmental sound classification from multi-feature data." *Applied Sciences* 12.12 (2022): 5988.

[14] Zhang, Zhichao, et al. "Deep convolutional neural network with mixup for environmental sound classification." *Chinese conference on pattern recognition and computer vision (PRCV)*. Cham: Springer International Publishing, 2018.

[15] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).

[16] Zamani, A. S., Hashim, A. H. A., Mohamed, S. S. I., & Alam, N. (2025). Optimized deep learning techniques to identify rumors and fake news in online social networks. Journal of Computational and Cognitive Engineering, 4(2), 142-150.

[17] Khalid Alkahtani, H., Mahmood, K., Khalid, M., Othman, M., Al Duhayyim, M., Osman, A. E., ... & Zamani, A. S. (2023). Optimal graph convolutional neural network-based ransomware detection for cybersecurity in IoT environment. Applied Sciences, 13(8), 5167.

[18] Zamani, A. S., Eltayeb, A. M., Alluhayb, A., Akhtar, M. M., Ayub, R., Abdelrahim, M. A. A., ... & Ahmad, N. (2025). Application of Machine learning in predicting cancer complications using longitudinal Data: A systematic review and Meta-Analysis. International Journal of Medical Informatics, 106217.

[19] Zamani, A. S., Jagadish, R. M., Kumar, B., Ghori, A. S., & Bhyratae, S. A. (2025). Perspectives of Machine Learning in the Convergence of Artificial Intelligence and Edge Computing. In Advances in AI for Cloud, Edge, and Mobile Computing Applications (pp. 189-211). Apple Academic Press.