Environmental Sound Classification using CNN

V Noothan
Department of AIE
University Amrita Vishwa Vidyapeetham
City, India
author1@example.com

M Thejeswara Reddy
Department of AIE
University Amrita Vishwa Vidyapeetham
City, India
author2@example.com

Abstract — Environmental Sound Classification (ESC) has become an essential component in the development of intelligent systems for applications such as urban monitoring, surveillance, and smart environments. In this study, we propose a deep learning-based framework that leverages Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction. The proposed architecture utilizes a Convolutional Neural Network (CNN) trained on the UrbanSound8K dataset, consisting of ten distinct environmental sound classes. By employing a consistent feature representation and efficient model design, the system achieves a high classification accuracy(93.13%) while maintaining a balanced precision and recall across all classes. The experimental results demonstrate the model's robustness and generalization capability, validating its effectiveness for practical ESC tasks. This work contributes a scalable and reliable approach for environmental sound recognition in real-world scenarios.

Index Terms — audio scene classification, DNN, RNN, CNN, i-vectors, late fusion

I. Introduction

In today's world, the ability of machines to understand and interpret environmental sounds plays a crucial role in the development of intelligent systems. Environmental Sound Classification (ESC) involves identifying sounds from the surrounding environment such as dog barks, sirens, car horns, and engine noises, which can carry vital information in both public and private settings. The context of this project lies in its application across various domains including surveillance systems, smart cities, assistive devices for the hearing-impaired, and home automation. Traditional methods relied heavily on manual feature extraction and classical machine learning algorithms, which often lacked the capacity to generalize across diverse sound environments. With the rise of deep learning, especially Convolutional Neural Networks (CNNs), sound classification has become more accurate and reliable. This project leverages MFCC features and a 1D CNN model to efficiently classify environmental sounds, providing a lightweight yet powerful solution. The significance of this research lies in its potential to contribute to real-time, intelligent audio recognition systems that can enhance safety, accessibility, and automation in our daily lives.

This work was supported by [Your Funding Agency], if applicable.

The unique aspect of our project is that we use a 1D Convolutional Neural Network (CNN) model to classify environmental sounds based on MFCC features. While many existing methods use 2D CNNs with spectrogram images, our approach is more lightweight and simpler. By using 1D convolutions directly on MFCCs, we reduce the complexity of the model and make it easier to train. This also makes our model suitable for real-time applications or devices with limited processing power. Our contribution is showing that even a simple model with the right features can achieve good performance in sound classification tasks. This makes our method efficient and practical for real-world use.

Our project has several practical advantages. First, it shows that environmental sounds can be classified accurately using only MFCC features, which are compact and easy to extract. This reduces the amount of data and processing required. Second, the 1D CNN model we used has fewer parameters, so it needs less memory and training time. This makes it ideal for real-time systems. Another advantage is that our approach can be easily extended or customized for different sound environments or datasets. The system can also help improve safety and awareness in public places by detecting sounds like alarms, sirens, or breaking glass. Overall, the project offers a balance between simplicity, speed, and performance.

The Environmental Sound Classification (ESC) system developed in this project has a wide range of practical applications in real-world scenarios. One major use is in smart surveillance systems, where automatic recognition of sounds like gunshots, screams, or glass breaking can help detect emergencies in real time. It is also useful in smart cities for monitoring urban noise pollution and ensuring regulatory compliance. In healthcare, ESC systems can assist in monitoring patients by detecting coughs, falls, or other abnormal sounds. Furthermore, it has applications in wildlife monitoring, where it can be used to track animal calls or detect illegal poaching activities in forests. Overall, the impact of this project lies in making machines more aware of their acoustic environment, improving safety, automation, and efficiency in various fields.

II. Literature Review

”Recent advancements in environmental sound classification (ESC) highlight the importance of feature extraction in enhancing classification accuracy. Previous studies have explored deep learning models, particularly Convolutional Neural Networks (CNN), for automatic feature learning from spectrogram images. The proposed method [1] demonstrates improved performance by combining deep features from CNN with KNN ensembles, achieving competitive accuracy on benchmark datasets such as DCASE-2017 ASC and UrbanSound8K

A recent study [2] proposed an environmental sound classification method based on multi-feature fusion, combining GFCC features derived from auditory perception and short-time energy features from the time domain. These features are fused to create a more comprehensive audio representation and are classified using an integrated network with a weighted voting mechanism. The method achieved an accuracy of 89.3% on the ESC-10 dataset.

A recent study [3] proposed an Urban Environmental Sound Classification model using multi-convolutional neural networks (CNNs), focusing on improving classification accuracy through model optimization. The research highlights the use of the Adam optimizer for better performance in urban soundscapes. This approach contributes to the growing body of work in applying deep learning techniques to environmental sound classification.

A 2022 study [4] proposed an environmental sound classification model leveraging CNN latent subspaces, focusing on feature extraction and optimizing training data. By employing a product Grassmann manifold approach, the model enhances the representation of CNN features, improving classification performance. This method contributes to the refinement of CNN-based models in environmental sound classification, enabling more accurate task analysis and classification in diverse acoustic environments.

A 2021 study [5] presented a deep convolutional neural network (CNN) for environmental sound classification, incorporating transfer learning to improve model performance. The approach utilized feature extraction techniques such as Mel-frequency cepstral coefficients (MFCC) and Log-Mel spectrograms, leveraging pre-trained Xception models. This method enhances classification accuracy on the ESC-50 dataset, demonstrating the effectiveness of transfer learning in environmental sound classification tasks.

A residual network with a dual attention mechanism was introduced for ESC, combining log-Mel spectrogram and MFCC features to enhance input representation. This model [6] applies channel and spatial attention to focus on critical time-frequency components, improving feature learning and recognition performance. It achieved a recognition rate of 95.42%, surpassing models using single features or no attention.

A hybrid expert model with an integrated attention mechanism was introduced to address the challenges of low signal-to-noise ratio and diverse sound types in ESC. This model [7] uses FBank method for feature extraction and replaces standard fully connected layers with multiple expert layers, which are dynamically selected by a routing layer to improve specialization and adaptability. Additionally, the attention mechanism enables the network to better capture important sound patterns by emphasizing salient features. On the UrbanSound8K dataset, the model achieved 96.0% accuracy, with precision, recall, and F1-score all reported at 0.961.

An environmental sound classification algorithm was proposed using adaptive data padding to effectively manage variable-length audio inputs. This technique [8] improves temporal alignment, allowing the model to focus on relevant features while minimizing the impact of padding noise. The approach enhances robustness and performance, particularly for short or uneven audio clips. On the UrbanSound8K dataset, the model achieved an accuracy of 96.45%, demonstrating strong classification capability.

A performance analysis was conducted on various acoustic feature aggregation strategies for environmental sound classification. The study [9] evaluated combinations of MFCC, log-Mel spectrogram, Chroma, Spectral Contrast, and Tonnetz features to determine optimal configurations for boosting classification performance. Through extensive experimentation, the fused feature set demonstrated superior effectiveness, achieving 85.6% accuracy on the ESC dataset and 93.4% on UrbanSound8K. The results highlight the importance of rich multi-feature representations in complex acoustic environments.

A CNN-based model was developed to classify urban sound events using short audio clips from the UrbanSound8K dataset. This approach [10] involves converting raw audio into log-Mel spectrograms using MFCCs, transforming them into image-like inputs for the CNN. The model architecture consists of multiple Conv2D and MaxPooling layers, followed by dense layers, and employs dropout to prevent overfitting. The system was trained for 250 epochs and achieved a classification accuracy of 91% on the test set, effectively demonstrating CNN's capability in audio-to-image classification for environmental sounds.

A stacked CNN-based approach was proposed for environmental sound classification using the UrbanSound8K dataset. The proposed method [11] utilizes MFCCs for feature extraction and explores the effects of various hyperparameters, including the number of convolutional layers and filter configurations. The optimized architecture includes three convolutional layers with increasing filters, followed by max pooling, dense layers, and dropout for regularization. The model achieved a testing accuracy of 93.8%, outperforming previous CNN-based methods. The study highlights the importance of model depth and feature hierarchy in improving ESC performance while suggesting future work in real-time and multi-source sound environments.

A CNN-based model was proposed for environmental sound classification using Mel-spectrograms as input features. This architecture [12] consists of six convolutional layers with LeakyReLU activation, followed by max pooling and dropout

to reduce overfitting. Three fully connected layers, including the final classification layer, are used to make predictions across 50 sound classes. The model incorporates MixUp data augmentation to improve generalization and was tested on a 2000-sample dataset collected from freesound.org. Compared to classic CNN models like AlexNet, ResNet-50, and VGG11, the proposed Mel-CNN achieved the highest accuracy of 81.75%, highlighting the effectiveness of deeper architectures and Mel-scale features for ESC.

A dual-model [13] approach was proposed for environmental sound classification using both 1-D and 2-D Convolutional Neural Networks (CNNs). The system processes audio signals through MFCC and Mel-spectrogram features, applying separate CNN architectures optimized for time-domain and time-frequency domain representations. The dataset used is UrbanSound8K, and the model incorporates preprocessing techniques like noise removal and log-frequency spectrogram transformation. The 1-D CNN achieved a high accuracy of 94.48%, while the 2-D CNN attained 90.07%, both outperforming traditional ANN and RNN-based models. The results validate the effectiveness of domain-specific convolutional strategies for accurate sound classification.

A deep learning model [14] was proposed for environmental sound classification using a Residual Network (ResNet) to address challenges in complex acoustic environments. The method extracts MFCC features from audio and processes them through a ResNet architecture enhanced with residual blocks and skip connections to avoid gradient vanishing and degradation issues. The network includes five residual blocks followed by pooling, LSTM, and fully connected layers, ending in a Softmax classifier. Experimental results on the UrbanSound8K dataset showed a classification accuracy of 90.4%, outperforming traditional shallow CNN models and demonstrating strong feature extraction and generalization capabilities.

A CNN-based ESC model [15] was proposed using novel bispectrum-derived representations to enhance classification accuracy. The study introduced multiple bispectrogram types—SBS, PBS, MPBS, DPBS1, and DPBS2—derived from third-order spectral analysis, each capturing different structural features of environmental sounds. The model was trained and evaluated on the UrbanSound8K dataset using a CNN with four convolutional layers and two fully connected layers. Experimental results showed that combining SBS with MPBS and DPBS1 improved accuracy from 71.5% to 72.6%, demonstrating the value of incorporating off-diagonal bispectral information. The approach highlights the potential of higher-order spectral features for robust sound classification in noisy or complex environments.

III. Methodology

For this project, we developed a deep learning-based system to classify environmental sounds into different categories. We used the UrbanSound8K dataset, which consists of 8732 audio clips spread across 10 classes such as dog barking, car horn, drilling, and more. The first step in our pipeline was to preprocess the audio files using the Librosa library in Python. Each audio file was loaded and resampled to ensure consistency across all samples. We then extracted Mel-Frequency Cepstral Coefficients (MFCCs), which are commonly used features in audio and speech recognition tasks. Specifically, we extracted 40 MFCCs from each audio clip, which effectively represent the timbral texture and frequency characteristics of the sound. Since audio clips vary in length, we averaged the MFCCs across the time dimension to obtain a fixed-size feature vector for every sample. This makes it easier to feed the data into a neural network without worrying about inconsistent input shapes.

After feature extraction, we used these vectors to train a 1D Convolutional Neural Network (CNN). The CNN architecture was designed to learn local patterns in the MFCC features by applying convolutional filters, followed by pooling layers to reduce dimensionality and capture the most important signals. We added dropout layers to prevent overfitting, and dense layers at the end of the network to perform the final classification. The model was trained using the Adam optimizer and categorical crossentropy loss, with techniques like early stopping and learning rate scheduling to improve performance and prevent overfitting. We also saved the best model using a model checkpoint during training. Finally, we tested our model on unseen audio samples and evaluated its performance using accuracy metrics. Overall, this approach allowed us to effectively classify environmental sounds with a good level of accuracy.

A. Feature Extraction (MFCC)

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot w[n] \cdot e^{-j2\pi kn/N}$$

It converts the time-domain audio signal into a frequency-domain representation for each short frame. $x[n]$ represents Discrete audio signal. $w[n]$ represents Window function (like Hamming). N represents frame length. k represents frequency bin index X[k] represents Frequency-domain value for bin k

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

Converts frequency scale into the Mel scale, matching human hearing, during MFCC computation. f represents frequency in Hz. f_{mel} represents Mel-scaled frequency

$$\text{MFCC}_k = \sum_{m=1}^{M} \log(E_m) \cdot \cos\left[\frac{\pi k(m-0.5)}{M}\right]$$

This calculates MFCCs from log-Mel energies, where MFCC_k represents the kth MFCC coefficient, E_m represents the energy of the Mel filterbank m, M represents the total number of Mel filters, and $\cos[\dots]$ represents the DCT basis function.

$$\mu_k = \frac{1}{T} \sum_{t=1}^{T} \mathrm{MFCC}_k(t)$$

It helps to get the average of MFCCs across time and turns variable-length sequences into fixed-size features, where T represents the total time frames, $\mathrm{MFCC}_k(t)$ represents the kth MFCC at time t, and μ_k represents the time-averaged MFCC.

B. CNN Feature Learning

$$y_i = \sigma \left(\sum_{j=1}^{K} w_j \cdot x_{i+j-1} + b \right)$$

It Extracts local temporal patterns from MFCC features using Convolutional 1D layers. x represents Input MFCC sequence (reshaped to shaped 40X1). w_j represents Learnable kernel weights. b represents bias term. σ represents activation function (ReLU). y_i output at position i.

$$y_i = \max(x_i, x_{i+1}, \ldots, x_{i+p-1})$$

It downsamples the feature map to reduce dimensionality using maxpooling layers. $X_i, ...$ represents convolved outputs in a window. p represents pool size. y_i represents max value in the pool window.

C. Loss Function Prediction

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

It converts final layer output into class probabilities. z_i represents output of last dense layer. C represents number of sound classes(10 in our dataset). $\hat{y}_i$ represents predicted probability for class i.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i)$$

It measures prediction error during training. y_i represents ground truth (one-hot encoded label). $\hat{y}_i$ represents predicted probability. $\mathcal{L}$ represents number of classes.

$$\text{Predicted Class} = \arg\max \hat{y}_i$$

It selects the class with highest predicted probability. $\hat{y}_i$ represents predicted class probabilities. $argmax$ returns index of the max value = predicted label.

This project is using UrbanSound8K dataset, which is a publicly available collection of urban audio recordings compiled for the task of environmental sound classification. The dataset consists of 8,732 labeled audio samples across 10 sound classes, including dog bark, air conditioner, gunshot, car horn, and drilling, among others.

Before feeding the audio data into our deep learning model, we performed several preprocessing steps. First, each audio file was loaded and resampled using the Librosa library to ensure consistency in sampling rate. We then extracted Mel-Frequency Cepstral Coefficients (MFCCs), which are known to be effective in capturing important features of non-speech audio. Specifically, we extracted 40 MFCCs from each file and computed the mean across the time axis to obtain a fixed-size feature vector. This ensured that all input samples had the same dimensionality, making them suitable for input into a Convolutional Neural Network (CNN). The resulting feature vectors were then labeled using one-hot encoding, and the dataset was split into training and testing sets using an 80/20 split.

IV. Results

The proposed model was evaluated using standard classification metrics, namely accuracy, sensitivity (recall), specificity, precision, and F1-score, computed both at the class level and as macro-averages. As shown in Table I and illustrated in the evaluation summary, the model achieved an overall accuracy of 93.13%, indicating strong general performance across the dataset. The macro-averaged sensitivity and specificity were 92.59% and 99.23%, respectively, demonstrating the model's excellent ability to correctly identify both positive and negative instances. Additionally, a macro precision of 93.35% and a macro F1-score of 92.93% further confirm the robustness of the classifier across all ten environmental sound categories.Per-class analysis reveals particularly high accuracy for classes such as engine_idling (99.54%), air_conditioner (99.48%), and siren (99.14%), with corresponding F1-scores exceeding 96%. While most classes show consistently strong performance, slightly lower recall values were observed for children_playing (87.32%) and gun_shot (87.36%), suggesting occasional misclassification in these more acoustically variable categories. Nevertheless, the balanced and high scores across metrics affirm the model's effectiveness and generalizability, especially in real-world urban sound environments.

The confusion matrix in Fig. 1 provides a detailed breakdown of the classification performance of the proposed model across all ten sound classes. It can be observed that most of the classes achieve high true positive rates, indicating strong model performance. For instance, the model correctly classified 193 out of 195 samples for the "air_conditioner" class, and 212 out of 215 for "engine_idling", reflecting the model's ability to learn and differentiate these sound patterns effectively. However, relatively higher misclassifications were seen in classes like "children_playing" and "street_music", where some samples were confused with similar ambient classes such as "dog_bark" and "siren", likely due to overlapping acoustic features in urban sound environments. Despite a few such confusions, the matrix reveals an overall strong diagonal dominance, suggesting that the model performs well in distinguishing among the sound classes. This analysis complements the quantitative evaluation metrics and demonstrates the robustness of our approach in real-world environmental sound classification tasks.

The graph presented in Fig. 2 illustrates the training and

TABLE I: Per-Class Evaluation Metrics

Class	Accuracy	Recall	Specificity	Precision	F1-score
Air conditioner	99.48	98.97	99.55	96.50	97.72
Car horn	99.26	90.11	99.76	95.35	92.66
Children playing	97.02	87.32	98.31	87.32	87.32
Dog bark	97.71	89.56	98.66	88.59	89.07
Drilling	98.40	91.58	99.29	94.39	92.96
Engine idling	99.54	98.15	99.74	98.15	98.15
Gun shot	99.20	87.36	99.82	96.20	91.57
Jack hammer	99.08	97.33	99.29	94.30	95.79
Siren	99.14	96.48	99.48	96.00	96.24
Street music	97.42	89.07	98.40	86.70	87.87
Average	93.13	92.59	99.23	93.35	92.93

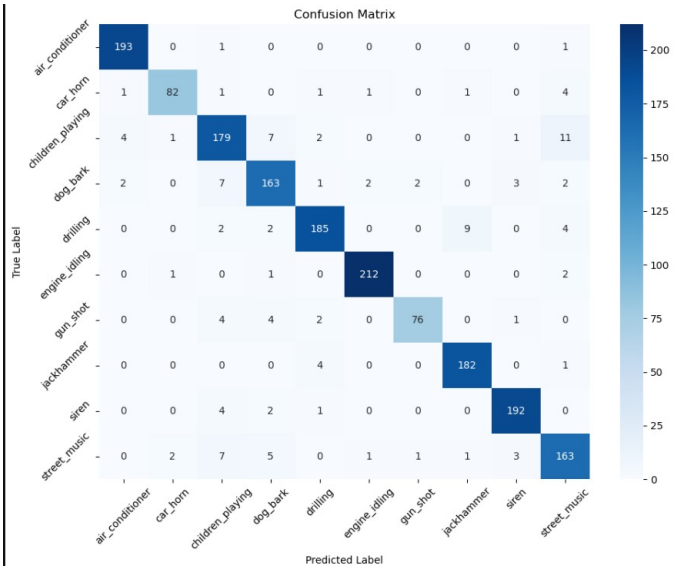

Fig. 1: Confusion matrix of the Proposed System

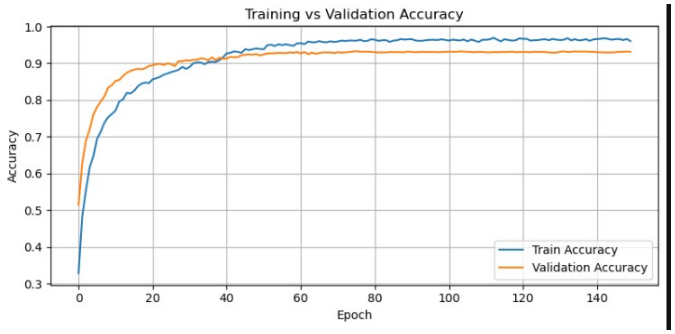

Fig. 2: Training vs validation accuracy of the Proposed System

Fig. 3: Training vs validation loss of the Proposed System

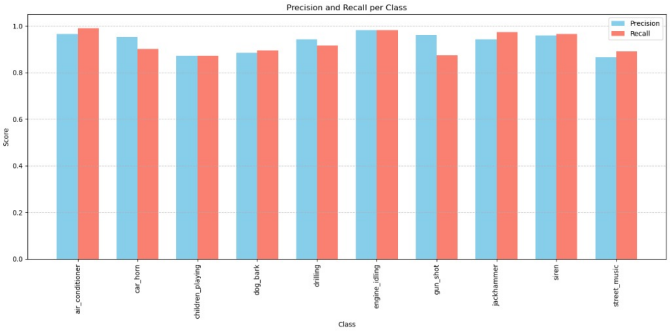

Fig. 4: Precision-recall Per-class of the Proposed System

validation accuracy over 150 epochs for the proposed Environmental Sound Classification model. Initially, both training and validation accuracy exhibit a steep increase, indicating effective learning during the early stages of training. Around the 40th epoch, the validation accuracy begins to plateau, stabilizing at approximately 92%, while the training accuracy continues to improve, eventually approaching nearly 98%. This divergence suggests that the model starts to overfit the training data beyond this point, although the gap remains

TABLE II: Comparison of environmental sound classification models based on various evaluation metrics

Reference	Accuracy (%)	Recall (%)	Specificity (%)	Precision (%)	F1-score (%)
[1]	86.70	86.726	98.523	86.826	86.754
[4]	85.01	-	-	-	-
[10]	91.00	-	-	-	-
[14]	90.40	-	-	-	-
[15]	72.44	-	-	-	-
Proposed method	93.13	92.59	99.23	93.35	92.93

relatively small, demonstrating that the model generalizes well to unseen data. The consistent performance on the validation set confirms the reliability of the feature extraction and model architecture employed in this work.

The Training vs Validation Loss graph in Fig. 3 shows how well the model learned over 150 epochs. In the beginning, both the training and validation loss drop quickly, which means the model is learning and improving. After about 40 epochs, the training loss keeps going down and gets very low, while the validation loss stays around 0.3. This small difference between the two losses shows that the model is not overfitting too much and is performing well on new, unseen data. Overall, the graph shows that the model was trained effectively and can make good predictions on data it hasn't seen before.

The bar graph in Fig. 4 displays the precision and recall scores for each sound class in the dataset. Overall, the model shows high performance across most classes, with both precision and recall scores generally above 0.85. Classes like air_conditioner, engine_idling, and siren have near-perfect scores, indicating that the model is highly accurate in identifying and correctly classifying these sounds. Slightly lower scores are observed for classes like children_playing and street_music, which may be due to the similarity of these sounds to other background noises, making them harder to distinguish. The balance between precision and recall across classes suggests that the model is not only accurate but also consistent in detecting true positives without too many false alarms.

V. Qualitative and Quantitative comparison

To assess the effectiveness of the proposed method, we compared its performance with several prominent models in the domain of Environmental Sound Classification (ESC). As presented in Table II, our model achieved an accuracy of 93.13%, along with a recall of 92.59%, specificity of 99.23%, precision of 93.35%, and F1-score of 92.93%. These results surpass or match the performance of notable existing methods such as the [1] model (86.70% accuracy) and the [10] approach (91.00% accuracy). While these models report high accuracy, they often lack comprehensive metric reporting or show limited performance consistency across other key indicators. In contrast, our model demonstrates robust and balanced performance across all evaluated metrics, indicating improved generalization and reliability. Additionally, earlier models like the [14] (90.40%) and [15] (72.44%) show lower accuracy, further highlighting the effectiveness of our approach.

VI. Conclusion

This work presents a deep learning-based method for environmental sound classification using MFCC features and a convolutional neural network architecture. The system was evaluated on the UrbanSound8K dataset, where it successfully classified audio samples into ten distinct environmental sound categories. The use of MFCCs provided a compact and informative representation of the audio signals, enabling the CNN to learn meaningful patterns. The model achieved high training and validation accuracy, with consistent precision and recall values across most classes. Visual analyses of loss and accuracy curves confirmed effective training with minimal overfitting. The results highlight the suitability of the proposed approach for real-world sound classification tasks. Future work can explore more advanced architectures, additional feature representations, and real-time deployment for practical ESC applications.

References

[1] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep cnn model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66 529–66 537, 2020.

[2] R. Li, B. Yin, Y. Cui, Z. Du, and K. Li, "Research on environmental sound classification algorithm based on multi-feature fusion," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9, 2020, pp. 522–526.

[3] K. Cai, B. Chen, and H. Zhang, "Research on urban environmental sound classification model based on multi-convolutional neural networks," in *2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2024, pp. 536–539.

[4] M. Mahyub, L. S. Souza, B. Batalo, and K. Fukui, "Environmental sound classification based on cnn latent subspaces," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[5] J. Lu, R. Ma, G. Liu, and Z. Qin, "Deep convolutional neural network with transfer learning for environmental sound classification," in *2021 International Conference on Computer, Control and Robotics (ICCCR)*, 2021, pp. 242–245.

[6] M. Sun, X. Sun, Z. Qiu, and L. Jia, "Research on environmental sound recognition method based on residual network with dual attention mechanism," in *2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 11, 2023, pp. 1814–1818.

[7] M. Zhou, F. Xia, and X. Zhao, "An acoustic model using mixture of experts for environmental sound classification," in *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, 2024, pp. 376–379.

[8] W. Qin and B. Yin, "Environmental sound classification algorithm based on adaptive data padding," in *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*, 2022, pp. 84–88.

[9] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Applied Acoustics*, vol. 158, p. 107050, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19302701

[10] M. Massoudi, S. Verma, and R. Jain, "Urban sound classification using cnn," in *2021 6th international conference on inventive computation technologies (icict)*. IEEE, 2021, pp. 583–589.

[11] S. Mittal, D. Patel, and M. Rahevar, "Elevating environmental sound classification with stacked convolutional neural networks," in *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 2024, pp. 1–6.

[12] Z. Gao, T. Liu, M. Zhu, J. Li, Y. Ning, and Z. Wang, "Environmental sound classification using cnn based on mel-spectogram," in *2023 2nd International Conference on Artificial Intelligence and Blockchain Technology (AIBT)*, 2023, pp. 41–45.

[13] L. S. Puspha Annabel, S. P. G, and T. V, "Environmental sound classification using 1-d and 2-d convolutional neural networks," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2023, pp. 1242–1247.

[14] C. Jiang, E. Li, and X. Yang, "Classification algorithm of environmental sound based on residual network," in *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, 2022, pp. 360–363.

[15] K. Hirata, S. Sakamoto, Y. Kobayashi, and S.-N. Suzuki, "Generation of effective bispectrogram for classification of environmental sound using convolutional neural network," in *2024 9th International Conference on Business and Industrial Research (ICBIR)*, 2024, pp. 0713–0717.