International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing Environmental Awareness of Deaf Individuals Using Deep Learning-Based Audio Classification

Aditya Kashyap[1], Manas Singh[2], Rajeshwar Kumar Dewangan[3]

Department of Computer Science and Engineering, Shri Shankaracharya Technical Campus, India[1]

Department of Computer Science and Engineering, Shri Shankaracharya Technical Campus, India[2]

Department of Computer Science and Engineering, Shri Shankaracharya Technical Campus, India[3]

ak158109@gmail.com[1] , 135790korba@gmail.com[2] , rajeshwarkd@gmail.com[3]

Abstract - ***Environmental sounds play a crucial role in human perception and situational awareness. However, individuals who are deaf or hard of hearing face significant challenges in detecting and interpreting such auditory cues, potentially compromising their safety and independence in everyday environments. This research presents a deep learning-based approach to real-time audio classification aimed at enhancing environmental awareness for deaf individuals. By leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the system is trained to recognize a wide range of environmental sounds—such as sirens, alarms, doorbells, approaching vehicles, and human speech patterns. The classified audio cues are then translated into intuitive visual or haptic feedback through wearable or mobile devices. Experimental results demonstrate high accuracy in sound classification and low latency in alert delivery, making the system suitable for real-world deployment. This approach holds significant potential to bridge the sensory gap for the deaf community and foster greater independence and situational awareness.***

Keywords - CNN, RNN, Audio Classification, Deep Learning, Machine Learning

I. Introduction

Environmental awareness is critical for navigating daily life safely and independently. While hearing individuals rely heavily on auditory cues—such as car horns, emergency sirens, barking dogs, or people shouting—those who are deaf or hard of hearing are often at a disadvantage in detecting and interpreting such sounds. This sensory gap can lead to increased vulnerability in dynamic environments like busy streets, public transportation hubs, or residential neighbourhoods where quick reactions to auditory cues are essential for safety.

To address this issue, we propose a deep learning-based audio classification system designed to enhance environmental perception for deaf individuals. Similar to how autonomous vehicles (AVs) require a robust perception system to safely navigate their surroundings, deaf individuals can benefit from a wearable or mobile-based assistive system that listens to the environment and interprets sounds in real time. For AVs, traditional sensors such as cameras, LiDAR, and radar provide visual and spatial data, but their effectiveness is limited by line-of-sight constraints and poor weather conditions. A similar limitation exists for deaf individuals relying solely on visual cues in complex and unpredictable environments.

In contrast, sound can provide valuable information beyond visual range—such as an approaching siren from an emergency vehicle, children playing behind a parked car, or a rapidly accelerating motorbike. Incorporating audio classification can help alert users to events that would otherwise go unnoticed. In the context of assistive technology, real-time audio recognition and categorization can serve as a digital “ear,” converting environmental sounds into intuitive visual or haptic feedback.

This paper presents a robust audio classification framework that utilizes a convolutional neural network (CNN) to identify and classify various urban sounds relevant to safety and situational awareness. Leveraging the UrbanSound8k dataset, which contains 8,732 annotated sound clips across 10 categories—including sirens, street music, children playing, and dog barking—the audio data is preprocessed into Mel Frequency Cepstral Coefficients (MFCCs) to be compatible with CNN input requirements. The proposed model demonstrates improved classification accuracy over existing approaches and lays the groundwork for integration into wearable or smartphone-based assistive devices for the deaf community.

By bringing the power of deep learning audio recognition into assistive contexts, this research aims to bridge the gap in environmental awareness for deaf individuals, enhancing both safety and quality of life.

II. Dataset Description

The UrbanSound8k dataset consists of 8,732 audio samples collected from real-world urban environments, categorized into ten distinct sound classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music (see Figure 1). Each sample is provided in ".wav" format, with a duration of less than five seconds, making them suitable for real-time processing in assistive applications. Figure 2 illustrates raw audio waveform samples from five representative classes: air conditioner, car horn, children playing, dog bark, and engine idling. Among these, only the dog bark sample is mono (single channel), while the others are stereo (dual channel), indicating varying audio complexities.

The dataset is accompanied by detailed metadata, including file names, source (www.freesound.org), file IDs, class IDs, occurrence IDs, and slice IDs, which help in tracing the origin and context of each audio clip. Each of the ten sound classes provides valuable information about the environmental context, such as potential hazards (e.g., sirens, gunshots), human activity (e.g., children playing), or nearby machinery (e.g., drilling, jackhammer).

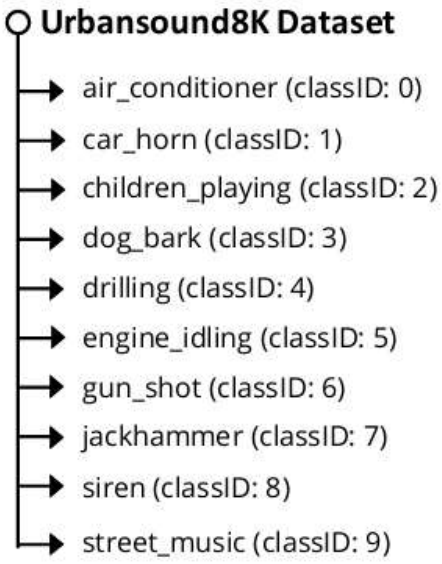

Fig 1. Urbansound8k dataset classes

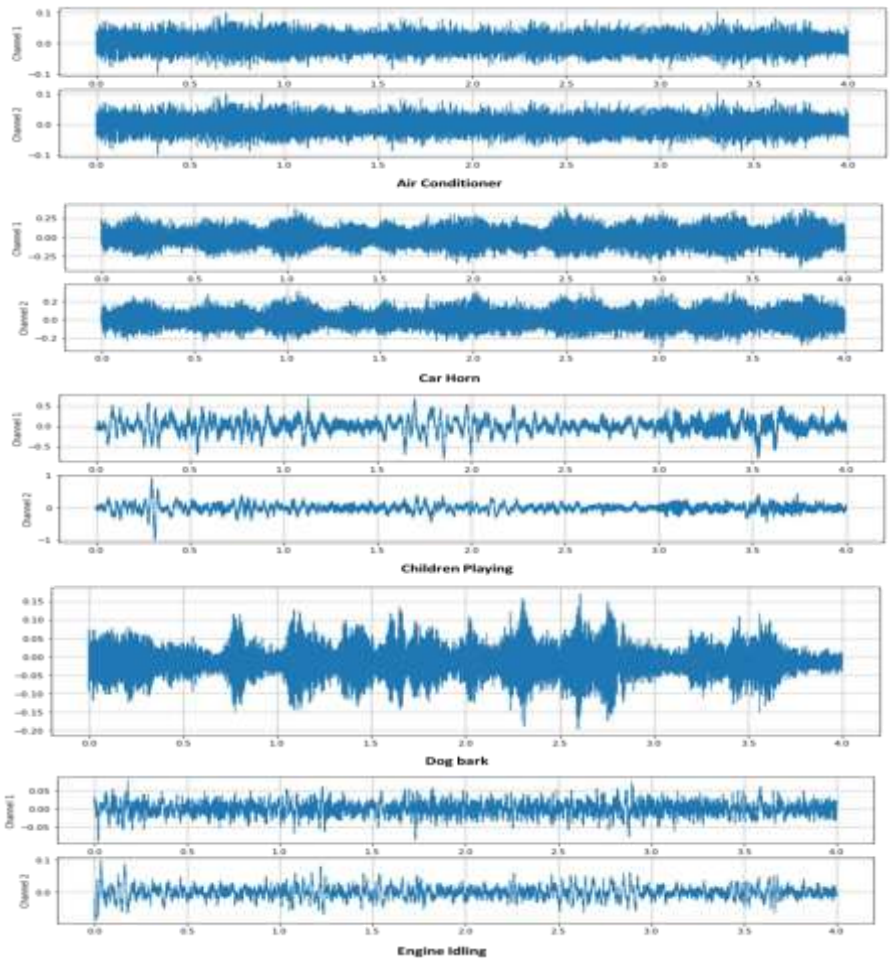

Fig 2. UrbanSound8k raw sample waveforms

III. Methodology

In this section, we presented the audio classification pipeline used in this study and compared several machine learning models in classifying the urban audio data. From existing literature, we found that convolution neural network (CNN) and long-short term memory (LSTM) are widely used for audio data classification. Therefore, we have used our audio classification pipeline to compare between our developed ML model and existing ML models. The audio classification pipeline, as shown in **Figure 3**, contains four basic blocks, i.e., dataset, splitting the dataset into training and validation, data preprocessing, and CNN audio classification framework.

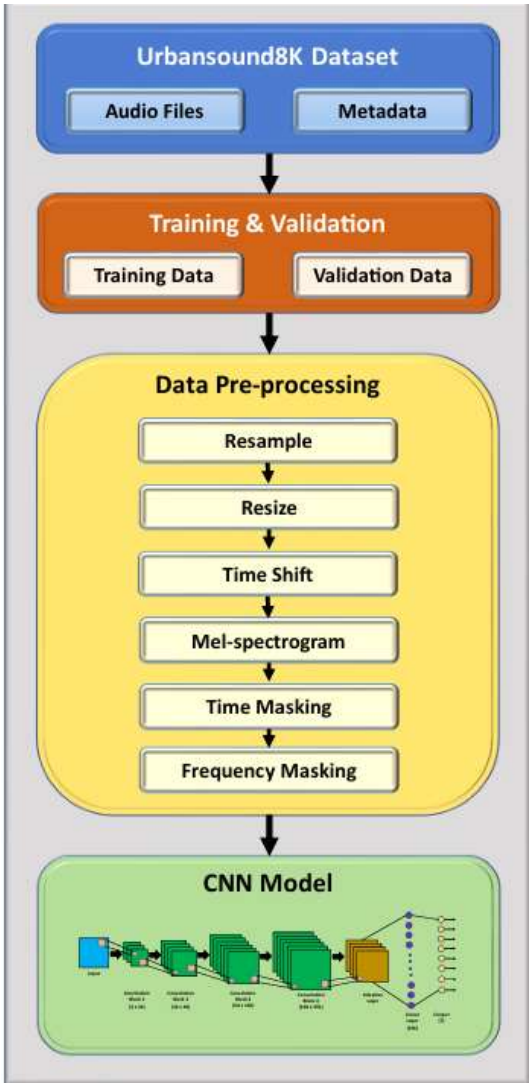

Fig 3. CNN-Based Audio Classification Pipeline

A. Data Preprocessing

The purpose of the data prepressing step is to create all the data into a unified format. In the Urbansound8k dataset, each sound is in a waveform format. In the first step, the data pre- processing block reads these waveforms formatted files. To ensure that all the samples are in stereo format, the audio is then resampled to 44,100 Hz and rechanneled to two channels. The resampled waveform is then padded or truncated to a specific fixed length. This process standardizes the data's length, sampling rate, and number of channels, ensuring that each feature has the same dimensions when extracted. The final step in the data preprocessing and standardization process is a time shift by a factor of 0.4, which is a data augmentation method equivalent to rotating or scaling images. This step ensures that the model can better generalize and make predictions for a larger variety of data for each given class.

B. Feature Extraction

Operating on the preprocessed waveform data, the program converts the audio into a Mel spectrogram, enabling it to be processed by an image classification neural network. Since the samples have the same dimensions, the Mel spectrograms are all the same size, eliminating the need for further standardization. After generating the Mel

spectrogram from the waveform, we perform data augmentation on the spectrograms. First, certain frequency ranges are masked. On a Mel spectrogram, this equates to horizontal bars over the plot, masking the values of randomly generated frequencies. A similar process occurs with time masking, which appears as vertical bars across specific time periods. Both masking steps mask up to 10% of their given dimension, ensuring that the data is still readable and that a machine learning model is capable of finding patterns consistently. The purpose of this data augmentation is to prevent overfitting and ensure that the model finds general patterns in the spectrograms, as opposed to overestimating the importance of specific training data features that will prevent generalization and cause overfitting. After data preprocessing, feature extraction, and data augmentation, each audio sample has two spectrograms of size 64x344. The two spectrograms result from the rechanneling to two-channel stereo specifications during the preprocessing step. **Figure 4** shows a sample visual representation of raw 4 seconds of audio data **(Figure 4(a))** and its corresponding spectrograms **(Figure 4(b)).** Finally, the data is normalized by **Equation 1**:

$$X = \frac{X - \mu}{\sigma} \qquad (1)$$

where, X is the input amplitude, μ is the mean of the input, and σ is the standard deviation of the input. This enables the model to ignore variable differences and classify samples solely based on a normalized representation of the spectrogram data.

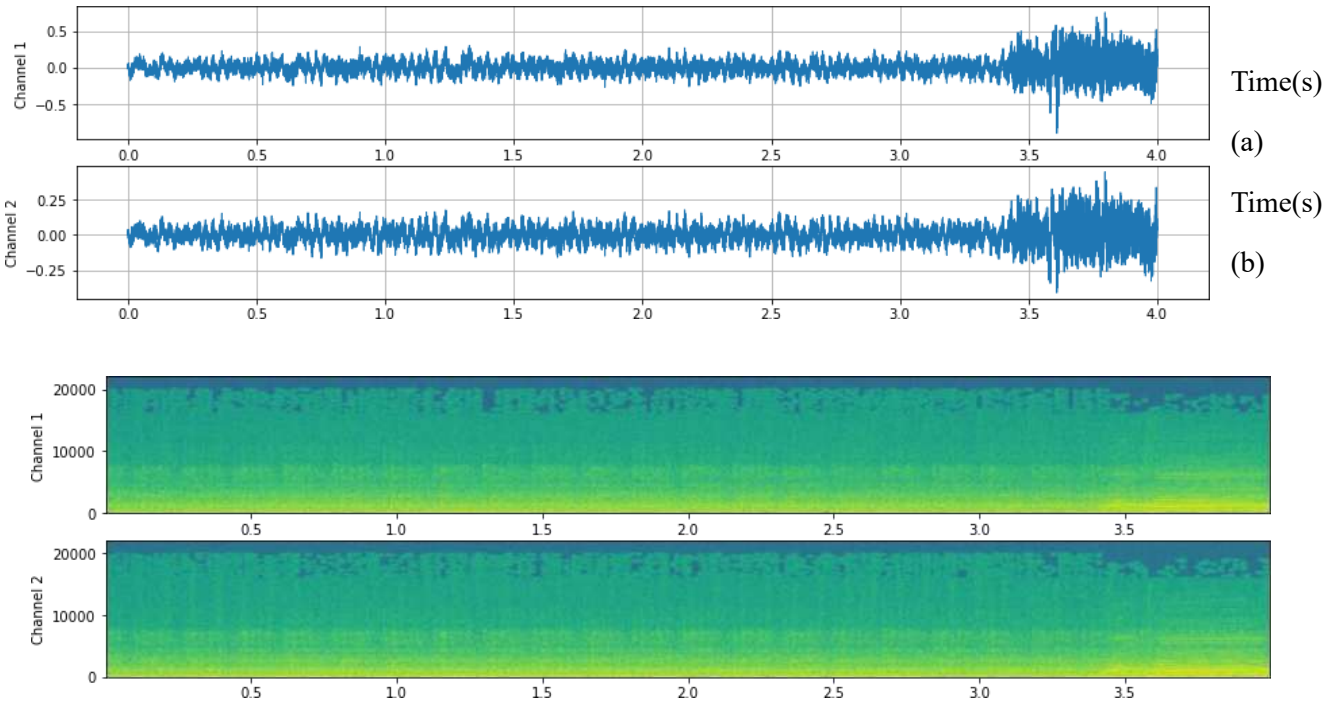

Fig 3. Visual display of audio data before (a) and after (b) converting to spectrogram

C. Audio Classification Model: Convolutional Neural Network (CNN)

After creating the normalized Mel spectrogram images, we use a convolution neural network (CNN) to classify the audio data. **Figure 5** presents a CNN architecture that utilizes four convolution blocks. The hyperparameters includes: Rectified Linear Unit (ReLU) as activation function, batch normalization, the Adam optimizer, and 2-D Average Pooling. This CNN model extracts the features through the convolution blocks, and classify the data into one of the 10 classes using fully connected (FC) layers. The model as a whole contains approximately 2.1 million trainable parameters. Table 1 provides the details of all the hyperparameters used in the construction of the convolution blocks. The convolution layers' hyperparameters are arranged to maintain a reasonable number of total parameters while enabling a powerful classification model. When tested with feature extraction filter sizes of 8, 16, 32, and 64, which is one of the primary hyperparameter

of the convolution layer, our accuracy rate suffered by 8%. Therefore, this larger model was used that decreases the training and classification efficiency without dramatically increasing the accuracy of the model. The padding and stride added further enhance the model's classification abilities by enabling the model to have a larger number of layers. Padding refers to addition of pixels (adding layers of zeros) to input images to preserve the border information of the image after using filter during a convolution operation. Whereas stride is a CNN filter parameter which detects the amount of movement over the input image. The ReLU activation function was the suitable choice for our model, which improved the classification accuracy further.

Finally, we normalized the audio inputs and include a batch normalization function after the activation function on each convolution block to reduce the training time.

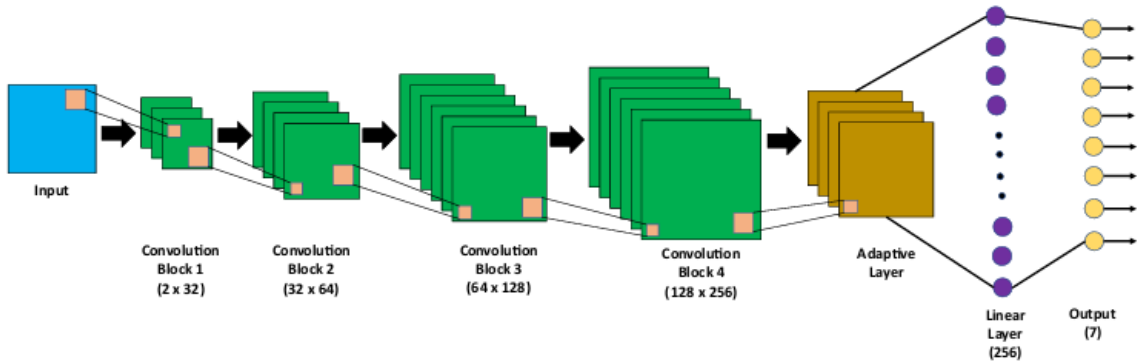

Fig 5. CNN Architecture used for audio classification

CONVOLUTION BLOCK	FILTER SIZE	KERNEL SIZE	PADDING	STRIDE
1	32	(3, 5)	(2,2)	(2,2)
2	64	(3, 5)	(2,2)	(1,1)
3	128	(5, 5)	(2,2)	(1,1)
4	256	(5, 5)	(2,2)	(1,1)

TABLE I Summary of CNN Layers

IV. Result and Discussions

A. Comparison with Existing Models

In this analysis, we consider all 10 classes that can be classified using ML models. First, we split the dataset into an 80/20 train-test split. After 100 epochs of training the model produced a training accuracy of 99% and a testing accuracy of 96.4%. The confusion matrix for this analysis is shown in Figure 8. Three out of the ten classes—i.e., air conditioner, car horn, and gunshot— attained a result of 99%, with children playing and engine idling having an accuracy of 98%. The two weakest classes were drilling at 94% and street music at 93%. As the confusion matrix demonstrates, air conditioning and car horn both have high accuracy rates of 99%. Each class is incorrectly selected over another class at a rate of 1%. Children playing has four instances of nontrivial rates of false positivity, with street

music being incorrectly identified as children playing at a rate of 3%. The dog bark has similar results to children playing, although its false negative rate is slightly higher. The drilling class and the jackhammer class are clearly confused at a high rate; however, the consequences of this confusion are largely insignificant in most cases. The engine idling class has high accuracy (98%), with a minor rate (1%) of confusion with drilling. The gunshot class is correctly identified with its accuracy of 99%, but it is falsely labeled as children playing with 1%. The siren class is falsely labeled as four other classes at a nontrivial rate, and street music was mistaken for a siren at a rate of 2%. Street music was falsely predicted as seven of the nine other classes, albeit at low rates for each one. Overall, the shortcomings of the model's predictive capabilities are most notably present in confusion between the drilling and jackhammer class, the false negative rate of the siren class despite its distinct sound, and the general low accuracy of the street music class.

As we compare the accuracy with the other existing ML models, our CNN model outperforms the presented model in the [1] and the [2] papers in a few key classes, most notably children playing and car horn. However, as shown in Analysis 1, this advantage of our model would be crucial in the context of autonomous vehicles due to the aforementioned situations. Table 2 shows the performance of our proposed model compared with that of the Two Stream CNN (TSCNN) [2] and that of the RACNN from [1]. The overall average does trail slightly in some of the sounds, such as engine idling, gun shoot siren, and street music, but for the other types of sounds our model outperforms in classifying/detecting the appropriate sound.

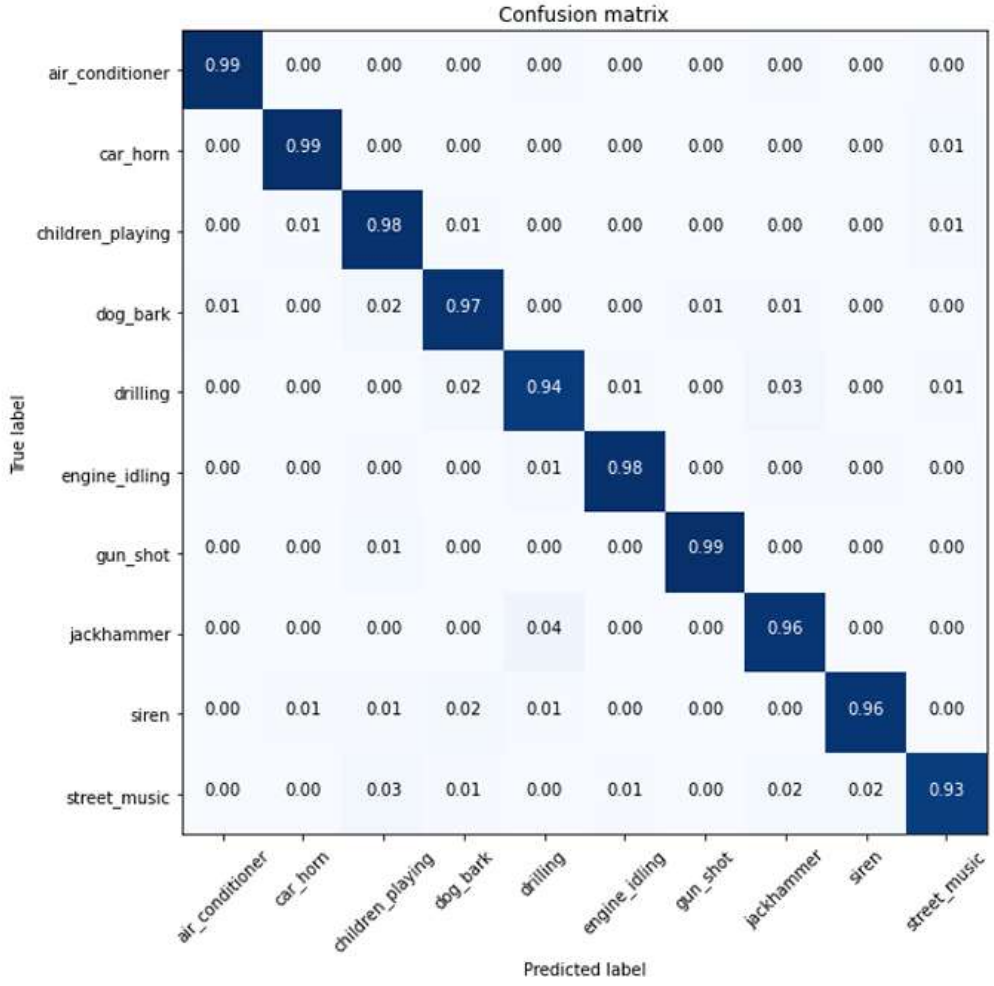

Fig 6. Confusion matrix of Comparison with Existing Models

Model	Air Conditioner	Car Horn	Children Playing	Dog Bark	Drilling	Engine	Gun Shoot	Jack Hammer	Siren	Street Music	Average
TSCNN [2]	.99	.94	.97	.95	.97	.99	.95	.97	.99	.97	.972
RACNN [1]	.98	.98	.95	.96	.98	1	1	.98	1	.95	.975
Our Model	.99	.99	.98	.97	.94	.98	.99	.96	.96	.93	.964

Table II Comparison with other models on individual classes in the UrbanSound8k dataset

V. Conclusion

In real-world scenarios where visual awareness may be compromised—such as during poor weather conditions, in crowded urban areas, or when visual cues are obstructed by trees, buildings, or construction—access to reliable auditory information becomes essential. For deaf individuals, the inability to perceive environmental sounds in such situations can pose serious safety risks and reduce situational awareness. The CNN model presented in this study, when integrated into a wearable or mobile device with appropriate sensory hardware (e.g., microphones), can serve as a powerful assistive tool to enhance environmental perception through real-time audio classification.

The evaluation results from our model demonstrate promising accuracy in recognizing a wide range of urban sounds, which can be crucial in everyday scenarios. For example, while walking near residential areas, the ability to detect and be alerted to the sound of children playing, barking dogs, or a nearby car horn can help prevent potential accidents. During storms or in noisy public spaces, recognizing subtle sounds such as an idling engine or a siren can offer critical cues that hearing individuals might take for granted.

The strong classification performance of our CNN—especially for sound classes with high safety relevance—suggests it is well-suited for deployment in assistive technologies aimed at the deaf and hard-of-hearing community. By providing timely, reliable feedback through visual or haptic alerts, the system has the potential to significantly increase safety, awareness, and independence in complex environments. These results support the model's future integration into deep learning-powered assistive solutions for inclusive, sound-aware living.

References

[1] Fang, Z., B. Yin, Z. Du, and X. Huang. Fast Environmental Sound Classification Based on Resource Adaptive Convolutional Neural Network. *Scientific Reports 2022 12:1*, Vol. 12,No. 1, 2022, pp. 1–18. https://doi.org/10.1038/s41598-022-10382-x.

[2] Su, Y., K. Zhang, J. Wang, and K. Madani. Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion. *Sensors 2019, Vol. 19, Page 1733*, Vol. 19, No. 7, 2019, p. 1733. https://doi.org/10.3390/S19071733.

[3] UrbanSound8K - Urban Sound Datasets. https://urbansounddataset.weebly.com/urbansound8k.html

[4] Salamon, J., C. Jacoby, and J. P. Bello. A Dataset and Taxonomy for Urban Sound Research. 2014

[5] Jangid, M., and K. Nagpal. Sound Classification Using Residual Convolutional Network. 12 Lecture Notes in Networks and Systems, Vol. 238, 2022, pp. 245–254. 13 https://doi.org/10.1007/978-981-16-2641-8_23/FIGURES/6.

[6] Su, Y., K. Zhang, J. Wang, and K. Madani. Environment Sound Classification Using a 26 Two-Stream CNN Based on Decision-Level Fusion. Sensors 2019, Vol. 19, Page 1733, 27 Vol. 19, No. 7, 2019, p. 1733. https://doi.org/10.3390/S19071733.

[7] Pelchat, N., & Gelowitz, C. M. (2020). Neural network music genre classification. Canadian Journal of Electrical and Computer Engineering, 43(3), 170-173.

[8] Nugroho, Devis Styo, Kusuma, Hendra, and Sardjono, T.. 2022. "Automatic Sound Alarm Classification Using Deep Learning For the Deaf and Hard of Hearing". https://doi.org/10.1109/ICSINTESA56431.2022.10041679

[9] Mou, Afsana and Milanova, M.. 2024. "Performance Analysis of Deep Learning Model-Compression Techniques for Audio Classification on Edge Devices". The Scientist. https://doi.org/10.3390/sci6020021

[10] Rezaul, Karim Mohammed, et al.. 2024. "Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models". Science and Information Organization. https://doi.org/10.14569/ijacsa.2024.0150704

[11] Zhang, Yixiao, Li, Baihua, Fang, Hui, and Meng, Qinggang. 2022. "Spectrogram Transformer for Audio Classification. https://doi.org/10.1109/ist55454.2022.9827729

[12] Wei, Shengyun, Zou, Shun, Liao, Feifan, and Lang, Weimin. 2020. "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification". IOP Publishing. https://doi.org/10.1088/1742-6596/1453/1/012085