South Asian Sounds: Audio Classification

Rajdeep Chatterjee[1,2]*, Pappu Bishwas[1], Sudip Chakrabarty[1], Tathagata Bandyopadhyay[2,3]
[1]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar 751024, India
[2]Amygdala AI, Bhubaneswar 751024, India
[3] Informatics, Technical University of Munich, Germany
Email: {cse.rajdeep, pappuovi8, sudipchakrabarty6, gata.tatha14}@gmail.com

2024 4th International Conference on Computer, Communication, Control & Information Technology (C3IT) | 979-8-3503-8912-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/C3IT60531.2024.10829485

***Abstract*—Sound classification is a significant task in the field of audio processing, with applications ranging from urban planning to noise pollution monitoring. We use Mel-Frequency Cepstral Coefficients (MFCCs) and a 1D Convolutional Neural Network (1D-CNN) model to try to solve the problem of putting urban sounds from Bangladesh into different groups. The proposed approach involves extracting MFCC features from audio recordings of urban environments in Bangladesh, which capture the spectral characteristics of the sounds. A 1D-CNN model then processes these MFCC features to classify the sounds into predefined categories like road-traffic noise, traditional Bengali songs, trains (railways), classroom noise, tanpura, etc. We also assess the performance of our approach on a dataset that has been collected and recorded from various locations in India and Bangladesh, as well as the well-known UrbanSounds8k dataset. Experimental results demonstrate the effectiveness of the proposed method, achieving high classification accuracy in distinguishing between different urban sound classes across both datasets. Our findings suggest that the combination of MFCCs and 1D-CNNs offers a robust solution for urban sound classification, with potential applications in urban planning, environmental monitoring, and smart city initiatives.**

***Index Terms*—Audio classification, CNN, MFCC, Sound recognition.**

I. Introduction

Sound classification has emerged as a vital area within the realm of audio processing, boasting diverse applications ranging from urban planning to noise pollution monitoring. In 1997, Sawhney and Maes, researchers at the Multimedia Laboratory at MIT, was the first to introduce and establish the pioneering concept of sound classification [1]–[3].

Sound classification has become increasingly significant in the field of audio processing, with applications spanning urban planning, environmental monitoring, and noise pollution mitigation. The study of sound classification traces its roots to early efforts in understanding and categorizing soundscapes within environments [4], [5].

The ability to discern and categorize sounds is crucial for gaining insights into the dynamics of urban settings, including human activities, transportation patterns, and environmental conditions. As such, the exploration of sound classification in urban contexts has garnered considerable attention from researchers and practitioners alike.

The evolution of sound classification parallels advancements in technology and computational techniques. Over time, researchers have devised various methods and algorithms for analyzing and classifying urban sounds, ranging from traditional signal processing techniques to more sophisticated machine learning approaches.

Early efforts in sound classification primarily focused on simple feature extraction methods such as spectrogram analysis and Fourier transforms. But since machine learning and deep learning came along, researchers are using methods like Mel-Frequency Cepstral Coefficients (MFCCs) and convolutional neural networks (CNNs) more and more to classify sounds more accurately and quickly [6]–[8].

In recent years, the availability of large-scale urban sound datasets, such as the UrbanSound8k dataset [9], has facilitated the development and evaluation of advanced urban sound classification models. These datasets enable researchers to train and test models on diverse urban sound samples, thereby improving the robustness and generalization capabilities of sound classification systems.

The integration of sound classification into smart city initiatives and environmental monitoring systems holds promise for enhancing livability and sustainability. By accurately classifying sounds, policymakers and planners can gain valuable insights into noise pollution levels, traffic patterns, and public safety concerns, thereby informing decision-making processes and improving the quality of life in different areas.

A. Contribution:

This paper presents a novel approach to South Asian Sounds (audio) classification using MFCCs and an one-dimensional convolutional neural network (1D-CNN) model. We leverage the rich spectral information captured by MFCCs and the discriminative power of 1D-CNNs to classify South Asian sounds into predefined categories. Through empirical evaluations on both self-collected datasets and benchmark datasets like UrbanSound8k, we demonstrate the effectiveness and practical utility of our proposed approach for South Asian sound classification in real-world settings.

B. Organization:

The paper contains seven sections. Section II provides a discussion of the related works. The details of the data preparation are in Section III and IV. Section V provides the description of the proposed model. We explain the results and performance in Section VI. Finally, the paper is concluded in Section VII.

Amygdala AI, is an international volunteer-run research group that advocates for *AI for a better tomorrow* http://amygdalaai.org/.

II. Related Works

Recent research has recognized the pressing issue of sound classification in urban areas, prompting investigations into effective detection and classification methods. Researchers have explored the application of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) Neural Networks [10]. In [11], authors have transformed the signals to images and classify with Resnet50v2 on urban sounds. Some models are employed using MFCCs (Mel-Frequency Cepstral Coefficients) as features and deep learning techniques. The aforementioned papers are the peer-reviewed research works with best performances to the best of our knowledge. The proposed model outsmarts these existing models and in addition, we introduce a new dataset for the community.

III. Data Preparation

A. South Asian Sounds Dataset

- **Audio Recording and Source:** The dataset used in this study comprises audio recordings obtained from mobile phones (Device: Google Pixel). The recordings capture a diverse range of acoustic characteristics in various environments and contexts. The audio recordings encompass five distinct classes, each corresponding to a specific category. It is part of a continuous project that contains more than ten types of audio specimens from most of the South Asian countries with common cultural values—India, Bangladesh, Sri Lanka, Afghanistan, etc. Based on the scope and performed research, only five categories are used in this study. The total size of the dataset is approximately 1.4 GB. Hereafter, we refer to this dataset as "**SAS-KIIT**". To assess the diversity of the SAS-KIIT dataset, we employed the t-SNE method and provided a corresponding visualization, which illustrates the distribution and variation of the data across different classes.

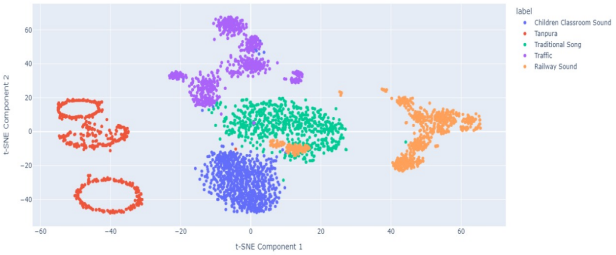

Fig. 1. TSNE

- **Sampling Rate Standardization:** User To facilitate uniform processing and analysis, all audio recordings, initially captured with varying sampling rates ranging from 32kHz to48kHz, have been resampled to a standardized sampling rate of 44.1kHz. This standardization step ensures consistency across the dataset, mitigating potential variations introduced by the different sampling rates inherent in mobile phone recordings.
- **File Format:** We stored the audio recordings in the WAV (Waveform Audio File Format) file format. We chose this format for its widespread compatibility and support for high-quality audio data. Additionally, WAV files preserve the integrity of the audio content, making them suitable for further analysis and processing.
- **Duration and Variability:** The duration of the audio recordings in the dataset varies, with some files exceeding 54 minutes in length. This variability reflects the natural diversity of mobile phone recordings, encompassing both short snippets and extended recordings. Such variability presents challenges and opportunities for developing robust classification algorithms capable of handling diverse audio lengths.
- **Class Distribution:** The dataset comprises recordings from five distinct classes, each representing different sound categories. We carefully selected these classes to cover a broad range of audio content commonly encountered in mobile phone recordings.

The class names are[1]:

- *Classroom Noise*
- *Tanpura*
- *Traditional Bengali Song*
- *Traffic (Roadways)*
- *Train (Railways)*

B. Audio Segmentation and Preparation

For each of the five classes, a total of 3750 segments have been generated by dividing the corresponding audio files into 750 segments each, with a fixed duration of 4 seconds, maintaining the original sampling rate. This segmentation approach aimed to create standardized audio samples while preserving the temporal characteristics of the original recordings. To ensure balanced representation and robust evaluation, the segments are randomly distributed across ten folders for 10-fold cross-validation purposes. Each folder contains almost an equal number of samples from all classes, facilitating comprehensive model evaluation. A metadata file documented essential information for each segmented sample, such as slice file name, slicing start and end times, class ID, class name, and folder ID. This metadata served as a reference for maintaining dataset integrity and facilitating reproducibility in future analyses.

IV. Sample Data Analysis and Visualization

We conducted visual inspection of audio samples from each class to identify similarities or patterns. Utilizing the librosa library, we loaded the sound files into arrays and visualized the audio waveforms. Subsequently, we employed a function to generate Mel spectrograms for further analysis. This function accepts parameters such as the file path, number of FFT points=2048, hop length=1024. Additionally, n-mels have been specified to control the number of Mel frequency bins. The resulting Mel spectrograms have been then converted

[1]SAS-KIIT Dataset samples: https://on.soundcloud.com/NWQBS

to decibel (dB) scale for visualization. Sample waveforms and their corresponding Mel spectrograms are presented below in Figs. 2-11.

A. *Data Visualization*

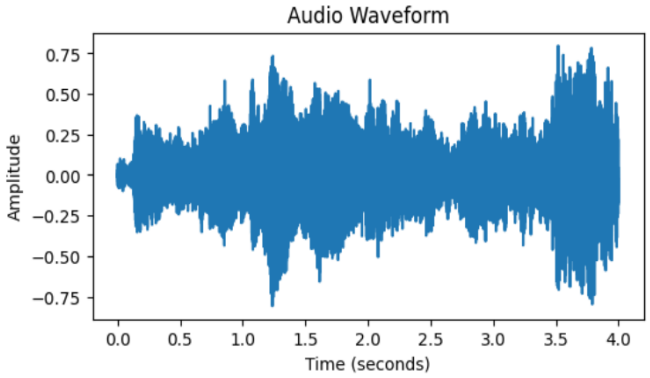

Fig. 2. Children Classroom Noise Waveform

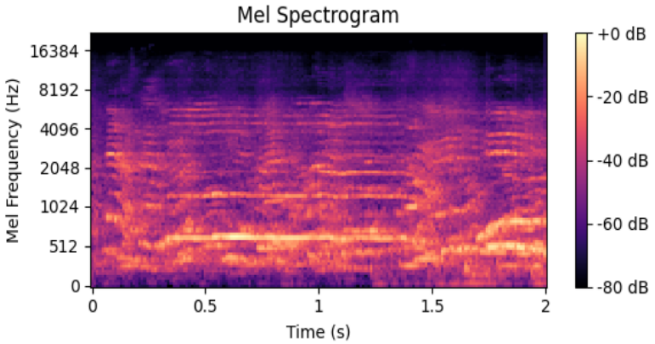

Fig. 3. Children Classroom Noise Mel-Spectrogram

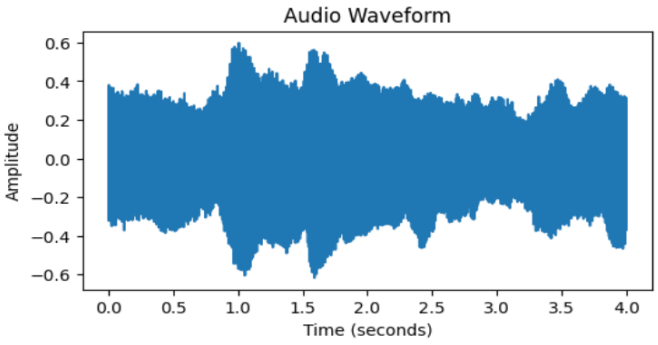

Fig. 4. Tanpura Waveform

B. *Feature Extraction*

In the process of feature extraction for audio data, Mel-frequency cepstral coefficients (MFCCs) play a pivotal role as they provide a valuable representation of the short-term power spectrum of a sound. Leveraging the librosa library, we employed the librosa.feature.mfcc function to compute

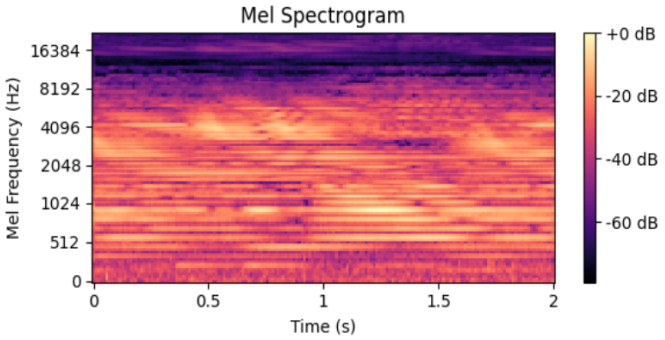

Fig. 5. Tanpura Mel-Spectrogram

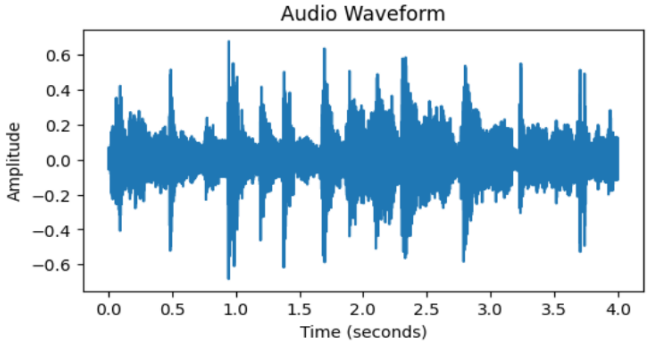

Fig. 6. Traditional Song Waveform

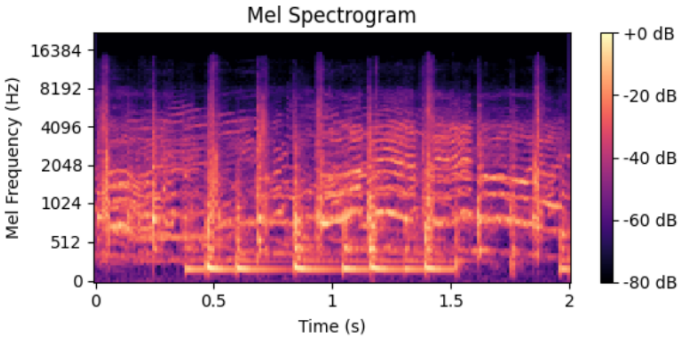

Fig. 7. Traditional Song Mel-Spectrogram

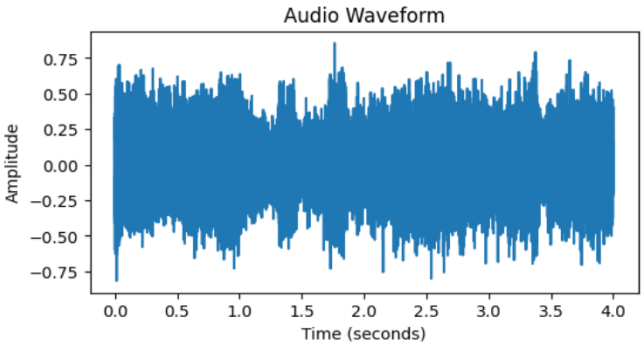

Fig. 8. Traffic Waveform

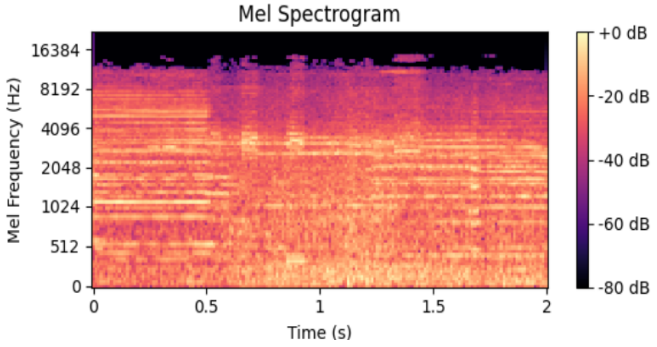

Fig. 9. Traffic Mel-Spectrogram

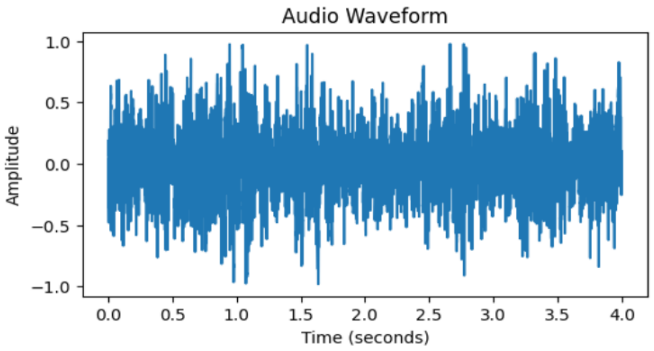

Fig. 10. Rail-Engine Waveform

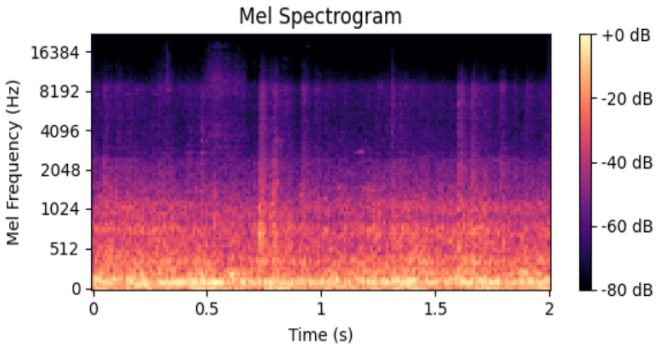

Fig. 11. Rail-Engine Mel-Spectrogram

MFCCs from the audio data (y), specifying the sample rate (sr) and the desired number of MFCCs to return (n-mfcc = 40) as parameters. Subsequently, the resulting MFCCs are examined, revealing a shape of $(40, 400)$, indicative of the computation of 40 MFCCs across 400 frames of the audio data. Furthermore, to enhance the effectiveness of feature representation, we applied standardization. Machine learning commonly employs this technique to standardize and compare data by transforming it to have a mean of zero and a standard deviation of one. Standardization using the StandardScaler involves the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- x = the original data
- μ = the mean of the data
- σ = the standard deviation of the data
- Z = the standardized data

The MFCCs have been computed and organized into a data frame (python data structure, DataFrame). The DataFrame encapsulates the extracted features under the 'Feature' column and stores the corresponding class labels in the 'class' column. We subsequently split the dataset into independent and dependent variables to facilitate machine learning analysis. The independent dataset, denoted as X, is a NumPy array containing the MFCC features, where each row represents a distinct observation and each column corresponds to an individual feature. Simultaneously, the dependent dataset, designated as y, comprises the class labels associated with each observation. This separation into independent features and dependent labels establishes a structured and coherent foundation for further exploration and modeling. These pre-processing steps are pivotal in preparing the audio data to feed into our model.

V. Model Building

Utilizing Keras, we construct a Convolutional Neural Network (1D-CNN) to address our classification task [12]–[14]. The model comprises sequential layers arranged in a structured manner to process input data effectively (see in Fig. 12).

A. *Proposed Model*

- **Input Layer:** The initial layer is a 1D Convolutional Layer with 64 filters, each having a kernel size of 3. The Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity into the network. This layer processes input sequences of shape (40, 400), where 40 represents the sequence length and 400 signifies the dimensionality of each feature.
- **Pooling Layer:** Following the convolutional layer, a MaxPooling1D layer is employed with a pool size of 2. This pooling operation helps in reducing the spatial dimensions of the feature maps, aiding computational efficiency and preventing overfitting.
- **Convolutional Layer:** Subsequently, another Conv1D layer is integrated with 128 filters and a kernel size of 3. Similar to the previous layer, ReLU activation is utilized to introduce non-linearity into the network. Another MaxPooling1D layer follows the second convolutional layer with a pool size of 2, further reducing the dimensionality of the feature maps.
- **Flattening Layer:** After the convolutional and pooling layers, a Flatten layer is employed to convert the 2D feature maps into a 1D vector, facilitating the transition to fully connected dense layers.
- **Dense Layers:** Two Dense layers are incorporated with 128 neurons each, utilizing ReLU activation. Additionally, a dropout rate of 20 percent is applied to regularize the network, aiding in preventing overfitting by randomly deactivating neurons during training.

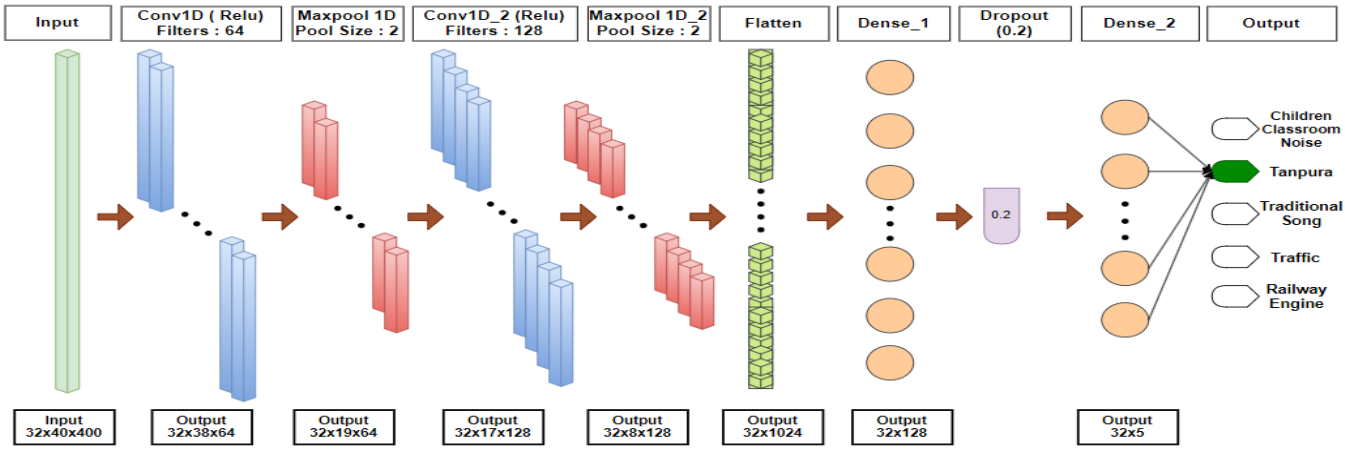

Fig. 12. Visualization of the proposed 1D-CNN model (SAS-CNN)

- **Output Layer:** The final layer is a Dense layer with 5 neurons, employing the softmax activation function. This configuration suits a multi-class classification problem with 5 distinct classes, ensuring that the output probabilities sum up to 1.

B. Compilation and Optimization:

The model is compiled using the Adam optimizer with a learning rate of 0.0001. Categorical cross-entropy is chosen as the loss function, suitable for multi-class classification tasks. Additionally, accuracy is monitored as the evaluation metric to gauge model performance during training. This architecture is meticulously designed to efficiently process sequential data, with measures in place to mitigate overfitting and enhance classification accuracy.

C. Training and Testing:

We conducted 10-fold cross-validation with a dataset of total 3750 samples. Each folder contains samples ranging from 361 to 391. For training, we use a step size of approximately 106, while validation has a step size of 13. The normal split consists of 3000 training samples and 750 test samples. During testing, the step size is set to 24, and for training, it was 94.

D. Early Stopping:

Early stopping with a large patience value of 5 is implemented to monitor the validation loss. This technique restores the model to its best weights when the validation loss fails to improve for consecutive epochs, preventing overfitting and promoting generalization.

This architecture is meticulously designed to efficiently process sequential data, with measures in place to mitigate overfitting and enhance classification accuracy.

VI. Results

We have implemented the proposed 1D-CNN model on a widely used audio classification dataset, that is, UrbanSounds8K. To establish that the proposed 1D-CNN model outperforms other similar audio classification models such as naive Multi Layered Perceptron (MLP), Deep Neural Networks (DNN), CNN, and Long Short-Term Memory (LSTM).

In [10], [11], the authors have performed rigorous experiments with feature extraction techniques and classifiers on the UrbanSounds8K dataset. Our suggested classifier model, which we'll call SAS-CNN from now on, does better than other SOTA models on the UrbanSounds8K dataset in hold-out (that is, train-test split), and it also does very well on 10-fold cross-validation (see Tables I and II).

Once we have established that the SAS-CNN is the top performer to the best of our knowledge on the UrbanSounds8K dataset. We examine the model's performance and robustness by employing it on our own South Asian Sounds (SAS-KIIT) dataset. The obtained results can be found in Table III.

TABLE I
ACCURACY (%) OBTAINED FROM DIFFERENT MODELS ON URBANSOUNDS8K DATASET

Models	UrbanSounds8k Accuracy (%)
ResNet50V2 [11]	90.70
MLP [11]	82.11
DNN100 [10]	90.90
CNN [10]	87.15
LSTM [10]	90.15
1D-CNN (proposed SAS-CNN)	93.50

TABLE II
ACCURACY (%) WITH 10-FOLD CROSS-VALIDATION OBTAINED FROM DIFFERENT MODELS ON URBANSOUNDS8K DATASET

Models	UrbanSounds8k Accuracy (%) 10-fold
1D-CNN (proposed SAS-CNN)	94.26±0.20

The confusion matrix, model's accuracy and loss graphs are shown in Figs. 13, 14, and 15. It can be concluded that the both

TABLE III
RESULTS OBTAINED FROM THE PROPOSED MODEL SAS-CNN ON SOUTH ASIAN SOUNDS DATASET

Models	South Asian Sounds (SAS-KIIT)	
	Accuracy (%)	Accuracy (%) 10-fold
1D-CNN (proposed SAS-CNN)	99.89	99.78±0.20

accuracy and loss curves are smooth. Therefore, the proposed scratch model is a stable classifier.

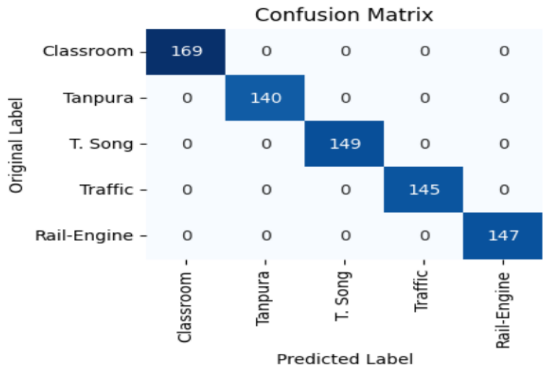

Fig. 13. Confusion Matrix obtained from our South Asian Sounds Dataset

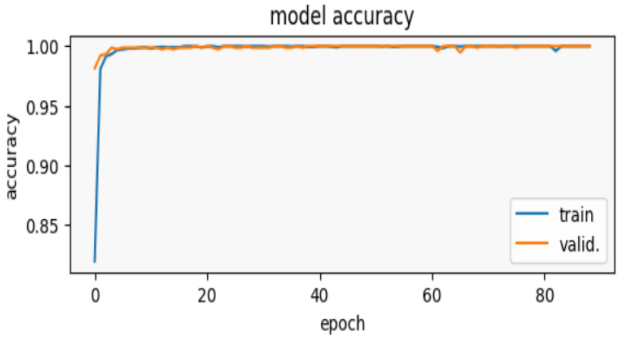

Fig. 14. Model Accuracy (%) obtained by 1D-CNN on our South Asian Sounds Dataset

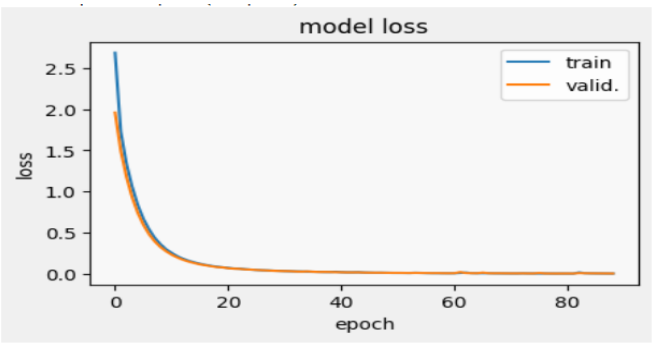

Fig. 15. Model Loss obtained by 1D-CNN on our South Asian Sounds Dataset

VII. CONCLUSION

Different biological and non-biological activities around us generate sounds. Audio (sound) classification is an important aspect of real-world activity monitoring, translation, and surveillance. Various types of sound classification still have room for improvement. The proposed framework, combining the extracted features and the SAS-CNN (1D-CNN) model, outperforms the existing SOTA models to the best of our knowledge.

In this paper, the authors have also introduced a new dataset, namely the South Asian Sounds dataset, with five different categories. The authors implemented the SAS-CNN model to classify the classes. It gives a very satisfactory performance both in holdout and 10-fold cross-validation.

We will expand the dataset with more South Asian-specific sounds in the future. Other feature extraction and selection techniques will be examined on the dataset.

REFERENCES

[1] N. Sawhney and P. Maes, "Situational awareness from environmental sounds," *Project Rep. for Pattie Maes*, pp. 1–7, 1997.

[2] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[3] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, 2023.

[4] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[5] U. Zölzer, *Digital audio signal processing*. John Wiley & Sons, 2022.

[6] S. Ali, S. Tanweer, S. S. Khalid, and N. Rao, "Mel frequency cepstral coefficient: a review," *ICIDSSD*, 2020.

[7] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.

[8] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, 2022.

[9] "Urbansound8k dataset," https://urbansounddataset.weebly.com/urbansound8k.html, July 2014, accessed: 2024-1-06.

[10] M. Bubashait and N. Hewahi, "Urban sound classification using dnn, cnn & lstm a comparative approach," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. IEEE, 2021, pp. 46–50.

[11] D. Xv and L. Yang, "Research on urban audio classification based on residual neural network," in *2021 International Conference on Computer Engineering and Application (ICCEA)*. IEEE, 2021, pp. 200–203.

[12] L. Eren, T. Ince, and S. Kiranyaz, "A generic intelligent bearing fault diagnosis system using compact adaptive 1d cnn classifier," *Journal of Signal Processing Systems*, vol. 91, pp. 179–189, 2019.

[13] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[14] A. A. Rahman and J. Angel Arul Jothi, "Classification of urbansound8k: A study using convolutional neural network and multiple data augmentation techniques," in *Soft Computing and its Engineering Applications: Second International Conference, icSoftComp 2020, Changa, Anand, India, December 11–12, 2020, Proceedings 2*. Springer, 2021, pp. 52–64.