Classification of Urban Sound using Custom 1D and 2D CNN Architectures

Shivam Sengar, Sunita Dhavale

Shivam Sengar, Defence Institute of Advanced Technology (DIAT), Pune, India (e-mail: shivamswngar4321@gmail.com)
Sunita Dhavale, Defence Institute of Advanced Technology (DIAT), Pune, India (e-mail: sunitadhavale@gmail.com)

ABSTRACT
In this work, we will study and analyze the performance of Convolutional Neural Networks (CNN) based techniques for sound classification which may help in case of urban surveillance applications. Recently, many researchers have used different feature extraction techniques along with different machine learning (ML) techniques to classify urban sound. Deep learning (DL) models like VGG-16, AlexNet, GoogleLeNet have also been explored in this direction effectively. In this research work, we proposed various custom designed 1D and 2D CNN models along with different extracted features like; Mel Frequency Cepstral Coefficients (MFCC), Chroma short-time Fourier transform (Chroma stft), Mel-spectrogram, and Spectral Contrast for urban sound classification. Detailed ablation studies were carried out with respect to different filter sizes, number of convolutional layers, and various activation functions to obtain state-of-the-art performances on the standard urban sound dataset. The experimental results show better performances with our custom CNN models compared to the classification models proposed in the existing recent literature works.

Keywords: Convolutional Neural Networks, Sound Classification, UrbanSound8K, Feature extraction, Deep Learning

1. INTRODUCTION

Sound classification tasks are useful in very diverse fields such as Security and Surveillance, Healthcare, Agriculture, Environmental Monitoring, Robotics, Education and Learning, Speech and Audio Processing, Industrial Applications, Smart Homes, Autonomous Vehicles, etc. With the expansion of cities, approximately 80% of the human population will live in cities by 2050. The cities will be well covered with various sensors and monitoring tools. With the growth of urbanization, the web of CCTV cameras is also expanding [1]. Most of the cameras are for video recording only. But very few of them are able to capture sounds. The video-only-camera can become useless if something happens which cannot be captured in a camera like a gunshot. So, we need cameras which can also record audio along the video for doing audio classification in real time. Suppose a person is shooting a gun and covering the gun with a bag, then the cameras would not recognize the gun shot. But if sound was recorded along with the video, then we can easily detect the gun shot. Secondly, distant things such as fighter jet cannot be easily detected by the image classification techniques. This will make the image classification technique confuse between a bird and the fighter jet. The detection of fighter jet like things make a lot of noise which is not possible for a bird to make. This way we can easily detect hidden or distant things which make a lot of noise in a very efficient way. Classification of Urban Sound will play a major role in the surveillance of our cities [2] in near future.

Any sound can be categorized into various class of sounds based on its acoustic features. Recently, many researchers explored various acoustic feature extraction techniques along with machine learning (ML) and Deep learning (DL) techniques for classification of urban sounds effectively [3]. In these works, many feature extraction techniques like Mel Frequency Cepstral Coefficients (MFCC) [4, 5], Log-Mel spectrogram [1, 4], Spectral Contrast, Chroma stft (short-time fourier transform) are employed to extract the important, detailed and prominent features from the sound data using Python Librosa library. After the feature extraction of the recordings, any ML technique such as K-Nearest Neighbour (KNN), random forest, Support Vector Machine (SVM), Principle Component Analysis (PCA) [1] or deep learning techniques such as DenseNet [12], VGG-16, VGG-19, AlexNet, etc. or combination of any of these can be employed for doing the audio classification task. Now-a-days, different types of models are being used for this task of audio classification. This includes long short-term memory (LSTM) networks [1], convolutional neural networks (CNNs) [6], artificial neural networks (ANNs), bidirectional LSTM (BILSTM) networks [5], Recurrent Neural Networks (RNN) [6] and many more.

Author in [7] used MFCC for feature extraction and a very basic custom CNN model to classify the environmental sounds with an accuracy of 64.5%. In 2019, Abdoli et. al. [8] suggested using 1D CNN as it can take direct input from the audio signal. It can also work with varying length input with the use of sliding window showing and accuracy of 89%. Mushtaq et. al. [9] proposed the aggregation of both Mel spectrogram and Log-Mel spectrogram features and DenseNet-161 model. In [10], the author discussed performances of various ML techniques like Hidden Markov Models (HMM), Decision Trees (DT), K-NN and SVM for sound data classifications. He also discussed DL techniques like, CNN, Multilayer Perceptron (MLP), Deep Neural Networks (DNN), and Recurrent Neural Networks (RNN) for sound data classifications. Author in [11] did a systematic literature review (SLR) to evaluate small dataset through data augmentation in order to increase the dataset. The author in ref. [12] used spectrogram images of the audio from ESC-50 and ESC-10 dataset to achieve an accuracy of 49% and 77% with CNN and achieved an accuracy of 56% on ESC-10 in tensor deep stacking network (TDSN).

In [13], the author first converted the audio signal into some suitable form using some signal representation techniques such as spectrograms, MFCC, wavelet decom position and linear predictive coding. After this, the author used five

different type of neural networks for the audio classification tasks, namely CNNs, Autoencoders, Transformers, Recurrent Neural Networks (RNNs) and Hybrid Models (CNN-RNN and CNN-Support Vector Machine). The author in [14] studied to identify both the recording device and the environment in which it is recorded. The author used 3 different environments (very quiet, quiet and noisy), 4 classes of recording devices and 136 speakers (68 male and female each), and 3600 recordings of sentences, words and continuous speech using CRNN (CNN+RNN). Sound classification using Mel spectrogram as feature extraction method and LSTM is explored in the study [15]. In [16], the author used MFCC as feature extraction method and a combination of three convolutional layer, three pooling layers, LSTM and flattening of the output and achieved an accuracy of 93.58% on UrbanSound8K dataset.

A combination of deep feature extraction, random subspaces K Nearest Neighbour (KNN) classifier and a custom CNN model is studied by the authors in [17]. The author in [18] used a combination of different feature extraction methods, consisting MFCC, Gammatone Frequency Cepstral Coefficients (GFCC), Constant Q-transform (CQT) and Chromagram in the deeper CNN (DCNN) achieving an accuracy of 97.52% on UrbanSound8K, 94.75% on ESC-10, and 87.45% on ESC-50. The author in [19] specially created manufacturing sound dataset (e.g. filing, hammering) and used a CNN, which receives log-Mel spectrograms of the sounds as input. The validation accuracy achieved was close to 100% displays an almost error-free performance; furthermore, the model converges very quickly to a stationary solution and also has very low validation loss. The study in [20] talks about determining the source of a weak sound, particularly in a busy or noisy surroundings and proposed a new attention-based context-aware neural network for weak environmental source classification. In [9], the model is based on simple log-power Short Time Fourier Transform (STFT) spectrograms and com bines them with several well-known approaches from the image domain (i.e., ResNet, Siamese-like networks and attention) achieving an accuracy of 97.0% (ESC-10), 91.5% (ESC-50) and 84.2% / 85.4% (US8K mono / stereo). Another study on Environment Sound Classification Task (ESC) is conducted by Iqbal et. al. [21] where the MFCC is used for feature extraction and we evaluate the use of CNN and multiple ML models to classify the sound signal using spectrograms of the sound spectrum. A study is conducted by Bensakhria et. al. [22] to detect domestic violence using classification of audio using 1D-CNN which outperformed classic machine learning-based models with 91.45% accuracy. Author in ref. [23] presents his own sound database recorded (NoisenseDB) in an urban environment and by using three DNN classifier he achieved 82%, 70% and 64% accuracy respectively.

Also, there is a large scarcity of labeled dataset in this domain. For e.g. Environ mental Sound Classification (ESC) dataset [24] consists of the sounds which can be easily available in our surroundings such as animal noise, traffic noise, human speech, etc. The sounds which are captured in a recording taken from an urban area are considered as Urban Sounds which includes sound samples like gun shot, siren, car horn, street music, dog bark, etc. The UrbanSound8K dataset contains most of these Urban Sounds [5]. Most of the above discussed works are either based on single feature analysis like MFCC [25] or GFCC [26], however considering various features may enhance feature extraction and analysis stages.

In this research work, we proposed various custom designed 1D and 2D CNN models along with different extracted features like; Mel Frequency Cepstral Coefficients (MFCC), Chroma short-time fourier transform (Chroma stft), Mel-spectrogram and Spectral Contrast for urban sound classification. A detailed ablation studies carried out by progressively expanding filter size in each consecutive convolution layer along with concatenation of various convolutional layers, using and various activation functions to get state of arts performances on standard urban sound dataset. Experimental results show better performances with our custom CNN models when compared to classification models proposed in existing recent literature works. The rest of the paper is organized as: Section 1 gives introduction with existing literature survey in detail. Section 2 explains proposed methodology. Section 3 gives details of experimental results and ablation studies performance comparisons. Section 4 concludes the findings.

2. METHODOLOGY

2.1 Dataset

The dataset used in this study is the UrbanSound8K dataset, which contains 8,732 labeled audio samples spanning 10 environmental sound classes. These include sounds such as dog barking, sirens, and street music, among others. The audio files are sampled at a 22,050 Hz sampling rate, and each file is assigned to one of the predefined classes. The class distribution of the dataset is shown in Figure 1.

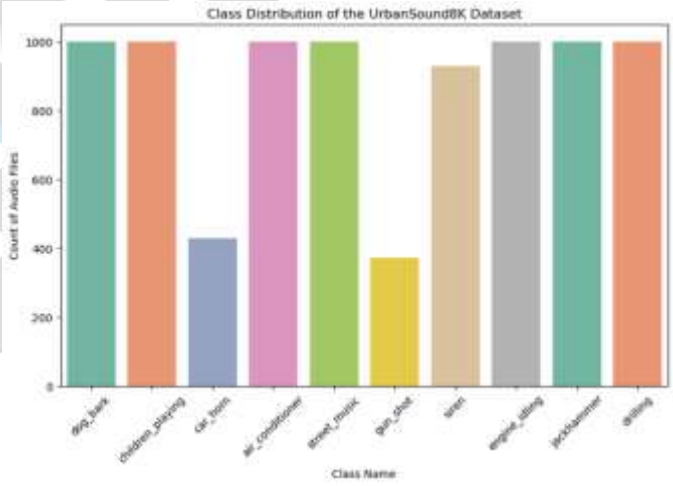

Figure 1: Class Distribution of the UrbanSound8K Dataset

2.2 Preprocessing

While processing audio segments, it is necessary to make length of all the audio segments of the same length. To resize them, we need to use padding which involves adding silence periods of zero values. To prepare the dataset for machine learning models, the audio samples were first pre-processed by extracting multiple

audio features. We utilized the following feature extraction techniques:

2.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC is commonly used feature in audio classification that captures the timbral texture of sound. These coefficients are derived from the Mel Spectrogram by applying a discrete cosine transform (DCT). The process starts by converting the audio signal into a Mel Spectrogram, then performing a logarithmic operation and a DCT to reduce the dimensionality. If $X(t,f)$ represents the spectrogram, $h_n(f)$ is the Mel filter bank, and M is the number of filters then,

$$MFCC_n(t) = \sum_{m=0}^{M-1} . \log\left(\sum_{f=0}^{F-1} |X(t,f) \cdot h_n(f)|^2\right)$$

represents MFCC coefficients.

Figure 2 and Figure 3 shows MFCC representations for sample audio files of Gun Shot and Siren respectively.

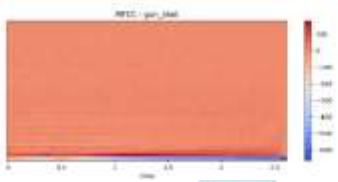

Figure 2: MFCC Representation of Sample: Gun Shot

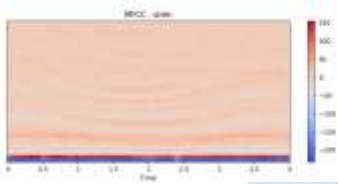

Figure 3: MFCC Representation of Sample: Siren

2.2.2 Chroma STFT Feature

Chroma STFT feature represents the twelve different pitch classes in a signal, useful for capturing harmonic content. If $X(t,f)$ is the spectrogram at time t and frequency f, N is the number of frequency bins, and w(k) is the weighting function, then

$$Chroma(t,f) = \sum_{k=0}^{N-1} X(t, f+k) \cdot w(k)$$

represents Chroma STFT. Figure 4 and Figure 5 shows Chroma Spectrograms for sample audio files of Gun Shot and Siren respectively.

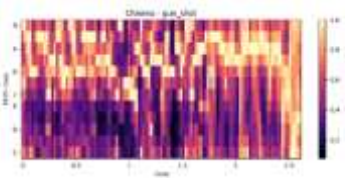

Figure 4: Chroma Representation of Sample: Gun Shot

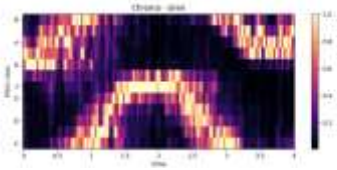

Figure 5: Chroma Representation of Sample: Siren

2.2.3 Mel Spectrogram

Mel spectrogram emphasizes lower frequencies and captures features useful for environmental sound classification. Mel Spectrogram is given as:

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

where f is the frequency in Hz and $M(f)$ is the corresponding Mel frequency. Figure 6 and Figure 7 shows Mel Spectrogram for sample audio files of Gun Shot and Siren respectively.

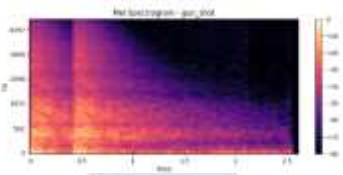

Figure 6: Mel Spectrogram for sample: Gun Shot

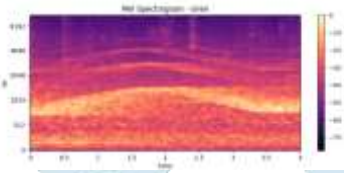

Figure 7: Mel Spectrogram for sample: Siren

2.2.4 Spectral Contrast

Spectral Contrast feature measures the difference in amplitude between peaks and valleys in a sound spectrum, which is particularly useful for identifying different sound textures. If P_i represents the power in the i-th subband and N is the total number of subbands, then

$$S_c = \sum_{i=1}^{N} . \log\left(\frac{P_{i-1}}{P_i}\right)$$

represents Spectral Contrast.

Figure 8 and Figure 9 shows Spectral Contrast Spectrograms for sample audio files of Gunshot and Siren, respectively.

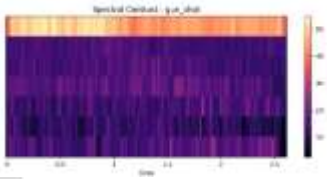

Figure 8: Spectral Contrast for sample audio: Gun Shot

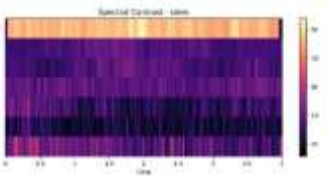

Figure 9: Spectral Contrast for sample audio: Siren

For each audio file, all above four different features (MFCC, Chroma stft, Mel spectrogram, and Spectral Contrast) were extracted, and the resulting feature arrays were padded or truncated to a fixed length of 174 time steps to ensure consistency across all samples. The feature arrays were then concatenated to form a single feature vector for each audio clip. This feature vector was used as the input to the subsequent models.

2.3 Model Architecture

We experimented with various Deep Learning (DL) models to classify sounds, with a focus on CNN architectures, which are well suited for extracting spatial patterns from data like spectrograms. We experimented with the performances of the following custom CNN models (Models 1-4) while comparing both Rectified Linear Unit (Relu) and Scaled Exponential Linear Unit (Selu) activation functions.

Model-1 as shown in Figure 10 is based on 1D CNN architecture, which is perfect for sequential data like raw audio signals since its convolutional layers handle the audio input in a one-dimensional manner. The SELU activation function, which normalizes activations across layers, is incorporated into the model to help with greater generalization and faster convergence. To ensure that the model performs effectively when applied to unseen data, dropout layers are used to prevent overfitting. Although 1-D CNN architectures provide less number of trainable parameters and lower computational complexity; 2-D CNN architectures not only exhibit high pre diction performances, but also are effective for capturing both spatial and temporal dependencies.

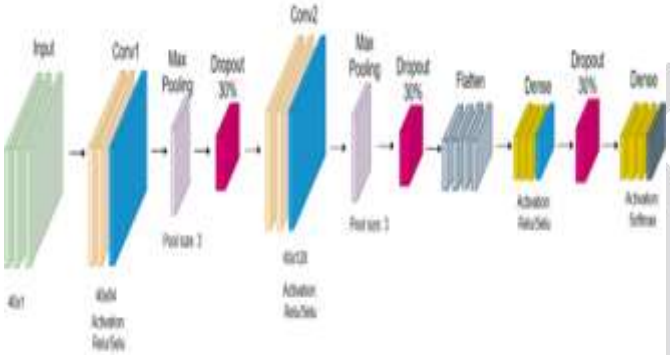

Figure 10: Model-1 based on 1-D CNN Architecture

In order to analyze spectrogram-like inputs, where both time and frequency dimen sions are important, Model-2 uses a 2-D CNN architecture as shown in Figure 11. To capture the spatial correlations between time and frequency components, the model makes use of 2-D convolutional layers. The SELU activation function is used to increase the convergence and stability of the model. Dropout layers are used to prevent over f itting. During detailed ablation studies, it is found that the accuracy of this Model is still less compared to previous 1-D architectures as only one convolutional layer with 3x3 filter sized is used which was not capable of capturing essential features of input data.

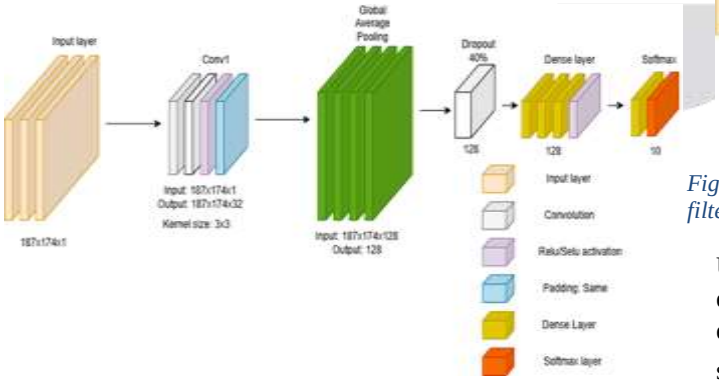

Figure 11: Model-2 based on 2D CNN Architecture

Hence, Model-3 based on 2D CNN architecture with two different convolutional layers, each using different filter sizes (3x3 and 5x5) as shown in Figure 12 is analyzed. Here, 8 3x3 filter being a small kernel size have a smaller receptive field and hence can extract small complex features. While, 5x5 being a mid-sized kernel has a large field view and hence can capture more global features. Finally, the output of both convolutional layers are concatenated to get good presentation of input features. During detailed ablation studies, it is found that the accuracy of this Model-3 with SELU (instead of RELU) activation function is enhanced more compared to Model-2.

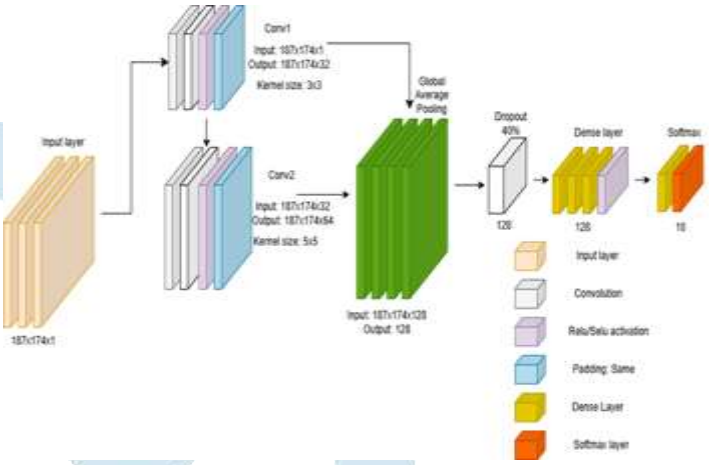

Figure 12: Model-3 based on 2D CNN Architecture with two different filter sizes

With a deeper network and bigger filter sizes, Model-3 goes beyond the 2D CNN technique to catch more intricate patterns in the input spectrograms as shown in Figure 13. Model-3 uses three different convolutional layers, each using different filter sizes (3x3, 5x5 and 7x7). Here, 7x7 being a large-sized kernel has a larger field view and hence can capture more global features apart from previously captured local features using 3x3 and 5x5 filters. During detailed ablation studies, it is found that the training procedure and model stability are improved by the continued assistance of the SELU (instead of RELU) activation function with self-normalization. In order to enhance its capacity for generalization and guarantee resilience against overfitting, the model additionally incorporates dropout layers along with Global average pooling (GAP) layer.

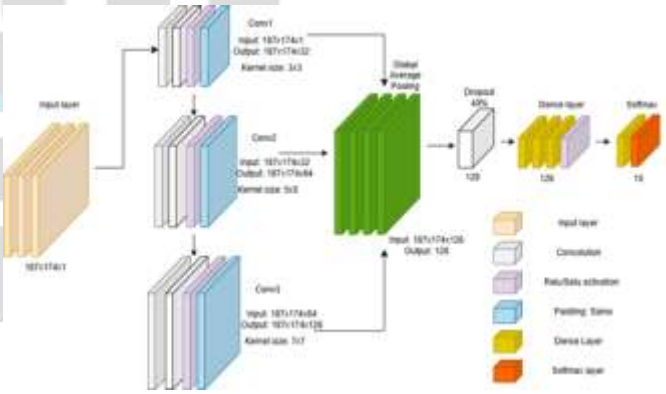

Figure 13: Model-4 based on 2D CNN Architecture with three different filter size

Using a more sophisticated 2D CNN architecture, Model-4 extracts a variety of information from the spectrograms by employing numerous convolutional layers with different filter

sizes. The model can efficiently learn both low-level and high-level characteristics because to this hybrid method. To ensure that the model works well without overfitting and with reduced complexity, dropout layers with GAP layer and SELU (instead of

RELU) activation function are used.

1D CNNs are perfect for smaller datasets because they effectively capture temporal patterns with fewer processing resources, which makes them suitable for sequential data such as raw audio signals. They have trouble, though, detecting spatial correlations between frequency and time, which is important for tasks like sound classification. On the other hand, 2D CNNs do better on these tasks because they are able to learn intricate patterns in both the time and frequency dimensions. But because they need more resources and are more computationally demanding, training becomes slower and more costly, particularly when dealing with huge datasets or high-resolution inputs.

The initial models consisted of several convolutional layers, followed by global pooling and fully connected layers. A common pattern across these models involved using 2D convolutional layers to process the spectrogram-like inputs. The first layer typically utilized smaller filters (e.g., 3x3) to extract local features, while deeper layers applied larger filters to capture more abstract patterns. MaxPooling or GlobalAveragePooling was used to reduce dimensionality, followed by Dropout layers to prevent overfitting and SELU (instead of RELU) activation function for getting more stability.

1. **Convolutional Layers:** Different combinations of convolutional layers with increasing filter sizes were experimented with, ranging from 32 filters with 3x3 kernels to 128 filters with 7x7 kernels.
2. **Pooling:** MaxPooling and GlobalAveragePooling layers were tested for reducing spatial dimensions and improving model generalization.
3. **Dropout:** A dropout rate of 0.4 was used to avoid overfitting.
4. **Hybrid Models:** Some of the enhanced models incorporated hybrid strategies, such as concatenating outputs from multiple convolutional layers with different filter sizes (e.g., 3x3, 5x5, 7x7) to create more diverse feature representations.
5. **Batch Size:** The batch was fixed to 32.

We also explored performance of SELU activation function, which are known for its self-normalizing properties, in place of ReLU to improve model convergence and stability.

2.4 Model Training

The models were trained using the Adam optimizer with sparse categorical cross entropy loss. The training process included early stopping to avoid overfitting, with the model saving the best weights based on validation loss performance. Additionally, TensorBoard was used for real-time visualization of training metrics such as loss and accuracy.

Data Augmentation: Instead of applying the data augmentation directly on raw audio features we used it in the feature extraction process to provide varied inputs by including different spectrogram representations.

Epochs and Batch Size: Model-1 is trained for up to 1000 epochs. Models 2-5 were trained for up to 50 epochs. All the models have a batch size of 32 samples. The training and validation sets were split using an 80-20 ratio, with stratification to ensure balanced representation of each class in both sets.

3. RESULTS

The performance of ten distinct models, developed using different architectures and feature combinations, was evaluated using the UrbanSound8K dataset. Metrics such as accuracy, precision, recall, and F1-score were computed to assess classification effectiveness. Below, we summarize the outcomes and insights derived from these experiments in Table 1:

Table 1: Accuracy, Precision, Recall, F1-score, and Number of Trainable Parameters of all the Models

Model	Accuracy	Precision	Recall	F1-score	Trainable Parameters
Model-1 (1D CNN)	0.91	0.91	0.91	0.91	75,530
Model-2 (2D CNN)	0.41	0.48	0.42	0.41	5,834
Model-3 (2D CNN)	0.82	0.83	0.83	0.83	65,290
Model-4 (2D CNN)	0.93	0.93	0.93	0.93	483,210
Model-1 (1D CNN) with Selu	0.91	0.91	0.91	0.91	75,530
Model-2 (2D CNN) with Selu	0.50	0.53	0.50	0.51	5,834
Model-3 (2D CNN) with Selu	0.85	0.85	0.85	0.85	65,290
Model-4 (2D CNN) with Selu	0.94	0.95	0.94	0.94	483,210

Table 2: Test Accuracy of all the Models

Model	Test Accuracy
Model-1 (1D CNN)	91%
Model-2 (2D CNN)	40.93%
Model-3 (2D CNN)	82.08%
Model-4 (2D CNN)	92.84%
Model-1 (1D CNN) with Selu	91%
Model-2 (2D CNN) with Selu	49.74%
Model-3 (2D CNN) with Selu	84.54%
Model-4 (2D CNN) with Selu	94.33%

From Table 1, we can infer that Model-4 performed best in overall precision. The Accuracy, Precision, Recall and F1-score of Model-1 based on 1D CNN models remain unchanged even when Selu activation function is introduced. While all the other models showed increased performance when Selu activation function is introduced. We got the highest test accuracy form Model-4 with Selu, as shown in Table 2.

3.1 Model Performance

Each model's accuracy and loss were recorded during training and validation phases. Models employing advanced features like SELU activation and Dropout exhibited improved generalization on the test data. Figure 14 and Figure 15 shows that the increased performance of all proposed Models when Selu is applied instead of Relu. Almost in every model using the Selu activation function, the accuracy is increased and the loss is decreased.

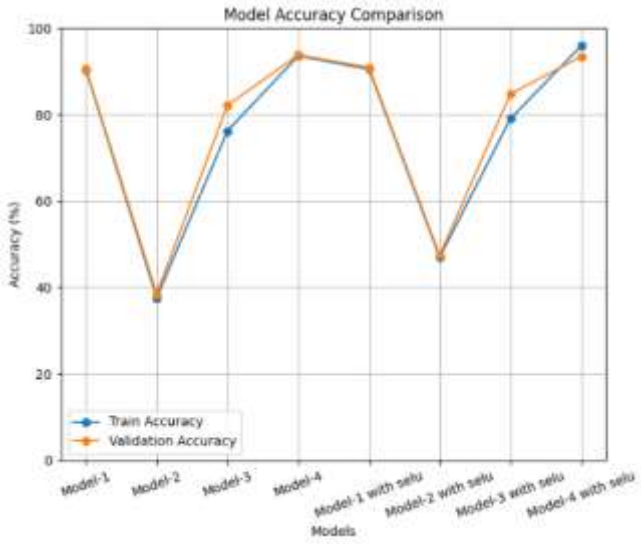

Figure 14: Graphical representation of Model Accuracy Comparison.

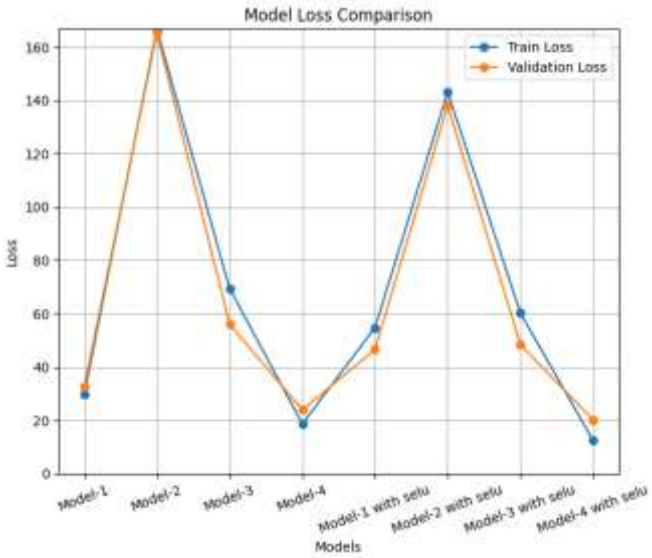

Figure 15: Graphical representation of Loss Comparison of the models.

From Figure 16 we can infer that the accuracy of Model-4 based on 2D CNN Architecture with three different filter sizes is increasing with the number of epochs. It shows the quality of our model during training. In Figure 17, we can see that both the training loss and validation loss are reducing with number of epochs.

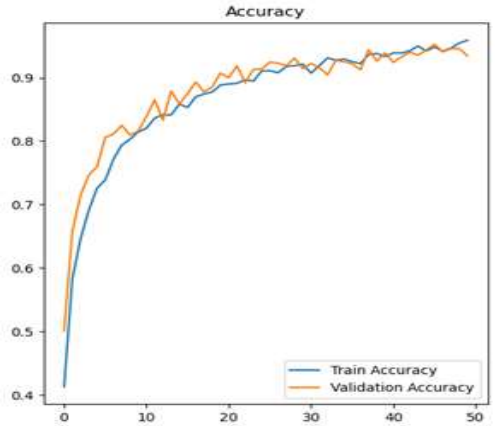

Figure 16: Accuracy curve of the best model Model-4 (2D CNN) with Selu out of all models.

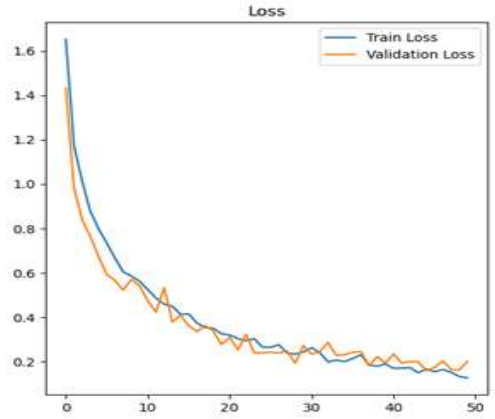

Figure 17: Loss curve of the best model Model-4 (2D CNN) with Selu out of all models.

In Figure 18, the confusion matrix of Model-4 with Selu is included which shows that most of the sounds are predicted correctly (TP), except dog bark(10 times) and street music(14 times) are sometimes considered as children playing. Gun shot has the highest percentage of TP in comparison to all the other classes. Gun shot is only one time considered as jackhammer, otherwise it is predicted correctly in all the other cases.

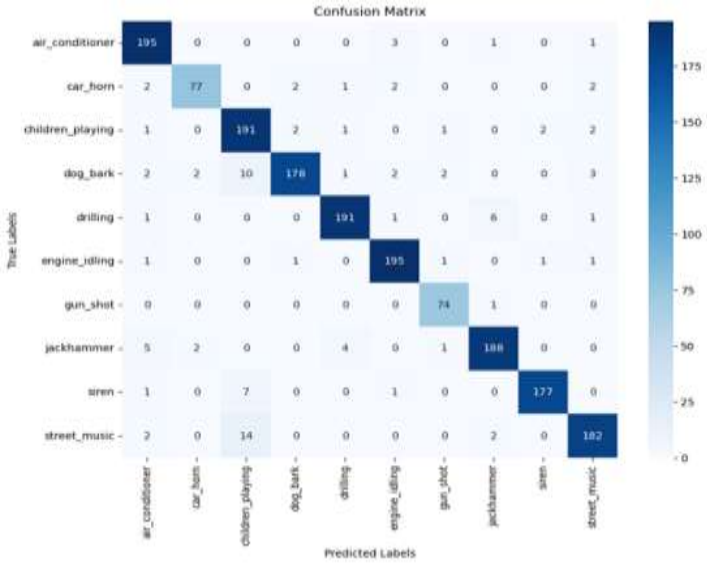

Figure 18: Confusion Matrix of our best model (Model-4 with Selu).

Results were visualized to compare the performance across models. Key findings include:

1. The enhanced architectures with Selu activation function consistently outperformed baseline models.

2. Training curves revealed that regularization strategies effectively mitigated overfit ting.

3.2 Comparison with Previously Studied models

Baseline models (Model-1 to Model-4) achieved moderate performance, highlighting the effectiveness of simple architectures for audio classification tasks. It is also observed that using Selu improved performance of Models-4, achieving the highest test accuracy of 94.33% on the test dataset. The performance of finally selected proposed model Model-4 with Selu is also compared with existing works on the same dataset. Table 3 shows the detailed comparison of the proposed Model-4 with Selu and 4 different selected features (MFCC, Mel-spectrogram, Spectral contrast, Chroma feature) with other existing works in terms of classification accuracy. It can be seen that proposed model outperformed the previous

studies (refer Table 3).

Table 3: Comparison of Different Models with Our Study.

Publishing Year	Preprocessing/Feature Extraction	Model	Accuracy	Remarks
2020 [27]	MFCC	Baseline	71%	Accuracy is low
2020 [27]	MFCC	DenseNet	81%	DenseNet network is better than baseline network
2020 [27]	GFCC	DenseNet	78.27%	GFCC performed inferior to MFCC
2020 [27]	MFCC	2-DenseNet	82.17%	Used improved form of DenseNet
2020 [27]	GFCC	2-DenseNet	79.57%	GFCC performed inferior to MFCC
2020 [27]	MFCC+ GFCC	2-DenseNet	82.75%	MFCC and GFCC combinedly extracted more features.
2020 [27]	[MFCC, GFCC]	D-2-DenseNet	84.83%	Used more advanced model than DenseNet and 2-Densenet
2021 [26]	Mel scale cepstral analysis (MEL)	CNN	87.15%	Tried a new feature extracton method.
2021 [28]	Mel scale cepstral analysis (MEL)	LSTM	90.15%	Using LSTM instead of CNN increased the accuracy.
2023 [29]	Data augmentation, MFCC	CNN	91%	Data augmentation helped the model to train for more number of epochs.
2024 [16]	MFCC	CRNN	93.58%	Accuracy can be increased by using aggregation of more feature extraction methods.
2024 (Our study: Model-4 with Selu)	MFCC, Mel-spectrogram, Spectral contrast, Chroma features	Custom CNN model with concatenation of Triple Convolution and Selu activation function	94.33%	High test accuracy, minimal training and validation loss.

4. Conclusion and Future Scope

In this study, various custom 1D and 2D CNN models are studied along with four different feature extraction methods. We used only one type of 1D CNN with relu and selu activation functions along with only MFCC features, both giving almost the same type of results. Whereas three types of 2D CNNs with progressively increasing f ilter size is used in each consecutive convolution along with relu and selu activation function with all four feature extraction methods, namely MFCC, Chroma stft, Mel spectrogram and Spectral contrast. During ablation studies, it was found that Model-4 with three 2D convolutional layers, each layer using different filter sizes and selu activation function, provides good test accuracy of 94.33%. Finally, from the above results, we can say that increasing the type of feature extraction methods and using concatenation of parallel convolutions having various filter sizes and Selu activation function can increase the performance of 2-D CNN architecture based DL frameworks for Urban Sound Data Classifications. In future, we will explore the concatenation of various features and novel CNN architectures in order to get more prominent features of audio files for classifications.

References

[1] Lezhenin, I., Bogach, N., Pyshkin, E.: Urban sound classification using long short-term memory neural network. In: 2019 Federated Conference on Com puter Science and Information Systems (FedCSIS), pp. 57–60 (2019). https: //doi.org/10.15439/2019F185

[2] Nogueira, A.F.R., Oliveira, H.S., Machado, J.J., Tavares, J.M.R.: Sound classification and processing of urban environments: A systematic literature review. Sensors 22(22), 8608 (2022) https://doi.org/10.3390/s22228608

[3] HONG, T.: 1-d and 2-d convolution neural network for bird sound detection. PhD thesis, Universiti Teknologi Malaysia (2020)

[4] Gupta, S., Srivastava, V., Kumar, D.: Environment sound classification using stacked features and convolutional neural network. In: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing, pp. 42–50 (2024). https://doi.org/10.1145/3675888.3676028

[5] Barua, S., Akter, T., Musa, M.A.S., Azim, M.A.: A deep learning approach for urban sound classification. International Journal of Computer Applications 975, 8887 https://doi.org/10.5120/ijca2023922991

[6] Xie, J., Colonna, J.G., Zhang, J.: Bioacoustic signal denoising: a review. Artificial Intelligence Review 54, 3575–3597 (2021) https://doi.org/10.1007/ s10462-020-09932-4

[7] Piczak, K.J.: Environmental sound classification with convolutional neural net works. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6 (2015). https://doi.org/10.1109/MLSP.2015. 7324337

[8] Abdoli, S., Cardinal, P., Koerich, A.L.: End-to-end environmental sound classification using a 1d convolutional neural network. Expert Systems with Applications 136, 252–263 (2019) https://doi.org/10.1016/j.eswa.2019.06.040

[9] Mushtaq, Z., Su, S.-F.: Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images. Symmetry 12(11), 1822 (2020) https://doi.org/10.3390/sym12111822

[10] Bansal, A., Garg, N.K.: Environmental sound classification: A descriptive review of the literature. Intelligent systems with applications 16, 200115 (2022) https: //doi.org/10.1016/j.iswa.2022.200115

[11] Abayomi-Alli, O.O., Damaˇseviˇcius, R., Qazi, A., Adedoyin-Olowe, M., Misra, S.: Data augmentation and deep learning methods in sound classification: A systematic review. Electronics 11(22), 3795 (2022) https://doi.org/10.3390/ electronics11223795

[12] Khamparia, A., Gupta, D., Nguyen, N.G., Khanna, A., Pandey, B., Tiwari, P.: Sound classification using convolutional neural network and tensor deep stacking network. IEEE Access 7, 7717–7727 (2019) https://doi.org/10.1109/ACCESS. 2018.2888882

[13] Zaman, K., Sah, M., Direkoglu, C., Unoki, M.: A survey of audio classification using deep learning. IEEE Access 11, 106620–106649 (2023) https://doi.org/10. 1109/ACCESS.2023.3318015

[14] Qamhan, M.A., Altaheri, H., Meftah, A.H., Muhammad, G., Alotaibi, Y.A.: Digital audio forensics: Microphone and environment classification using deep learning. IEEE Access 9, 62719–62733 (2021) https://doi.org/10.1109/ACCESS.2021.3073786

[15] Tyagi, S., Aggarwal, K., Kumar, D., Garg, S., et al.: Urban sound classification for audio analysis using long short term memory. NEU Journal for Artificial Intelligence and Internet of Things 1(1), 1–11 (2023)

[16] Bansal, A., Garg, N.K.: Robust technique for environmental sound classification using convolutional recurrent neural network. Multimedia Tools and Applications 83(18), 54755–54772 (2024) https://doi.org/10.1007/s11042-023-17066-2 17

[17] Demir, F., Abdullah, D.A., Sengur, A.: A new deep cnn model for environmental sound classification. IEEE Access 8, 66529–66537 (2020) https://doi.org/10.1109/ACCESS.2020.2984903

[18] Sharma, J., Granmo, O.-C., Goodwin, M.: Environment sound classification using multiple feature channels and attention based deep convolutional neural network. Interspeech (2020) https://doi.org/10.21437/Interspeech.2020-1303

[19] Fink, L.: Sound is context: Acoustic work step classification using deep learning. PhD thesis, Technische Universit¨ ”at Wien (2022)

[20] Presannakumar, K., Mohamed, A.: Source identification of weak audio signals using attention based convolutional neural network. Applied Intelligence 53(22), 27044–27059 (2023) https://doi.org/10.1007/s10489-023-04973-y

[21] Iqbal, K., Din, A.U., Alim, A., Sadiq, U.B., Ahmed, S.: Performance evaluation of environmental sound classification: A machine learning stacking and multi-criteria metrics based approach. Quaid-e-Awam University Research Journal of Engineering Science and Technology 21(1), 77–86 (2023) https://doi.org/10.52584/QRJ. 2101.10

[22] Bensakhria, A.: Detecting domestic violence incidents using audio monitoring and deep learning techniques https://doi.org/10.13140/RG.2.2.36128.97280/1

[23] Diez, I., Saratxaga, I., Salegi, U., Navas, E., Hernaez, I.: Noisensedb: An urban sound event database to develop neural classification systems for noise monitoring applications. Applied Sciences 13(16), 9358 (2023) https://doi.org/10.3390/app13169358

[24] Das, J.K., Ghosh, A., Pal, A.K., Dutta, S., Chakrabarty, A.: Urban sound classification using convolutional neural network and long short term memory based on multiple features. In: 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), pp. 1–9 (2020). https://doi.org/10.1109/ICDS50568.2020.9268723

[25] Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal processing letters 24(3), 279–283 (2017) https://doi.org/10.1109/LSP.2017.2657381

[26] Madhu, A., K, S.: Envgan: a gan-based augmentation to improve environmental sound classification. Artificial intelligence review 55(8), 6301–6320 (2022) https://doi.org/10.1007/s10462-022-10153-0

[27] Huang, Z., Liu, C., Fei, H., Li, W., Yu, J., Cao, Y.: Urban sound classification based on 2-order dense convolutional network using dual features. Applied Acoustics 164, 107243 (2020) https://doi.org/10.1016/j.apacoust.2020.107243 18

[28] Bubashait, M., Hewahi, N.: Urban sound classification using dnn, cnn lstm a comparative approach. In: 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 46–50 (2021). https://doi.org/10.1109/3ICT53449.2021.9581339

[29] Chu, H.-C., Zhang, Y.-L., Chiang, H.-C.: A cnn sound classification mechanism using data augmentation. Sensors 23(15), 6972 (2023) https://doi.org/10.3390/ s23156972

AUTHORS

Shivam Sengar received his BTech degree in Computer Science & Engineering from the Khwaja Moinuddin Chishti Language University, Lucknow, India in 2023. He is currently pursuing his MTech in Computer Science and Engineering, specialized in Artificial Intelligence from the Defence Institute of Advanced Technology, Pune, India. His areas of interest are Deep Learning, Machine Learning, Computer Vision, and Data Analytics. E-mail: shivamswngar4321@gmail.com

Sunita Vikrant Dhavale received her ME in Computer Science from Pune University, India in 2009 and Ph.D. degrees in Computer Science from Defence Institute of Advanced Technology, Pune, India in 2015. She is presently associated with the Defence Institute of Advanced Technology, Pune, as an Associate Professor in the Department of Computer Engineering. Her areas of interest are computer vision, deep learning, and cybersecurity. Email: sunitadhavale@gmail.com