I.J. Modern Education and Computer Science, **2020, 5, 41-54**
Published Online October 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2020.05.04

Automatic Environmental Sound Recognition (AESR) Using Convolutional Neural Network

Md. Rayhan Ahmed
Department of Computer Science and Engineering, Stamford University Bangladesh, Dhaka, Bangladesh
Email: rayhansimanto@gmail.com

Towhidul Islam Robin
Department of Computer Science and Engineering, Stamford University Bangladesh, Dhaka, Bangladesh
Email: towhid.austcse@gmail.com

Ashfaq Ali Shafin
Department of Computer Science and Engineering, Stamford University Bangladesh, Dhaka, Bangladesh
Email: shafinashfaqali21@gmail.com

Received: 25 March 2020; Accepted: 08 May 2020; Published: 08 October 2020

Abstract: Automatic Environmental Sound Recognition (AESR) is an essential topic in modern research in the field of pattern recognition. We can convert a short audio file of a sound event into a spectrogram image and feed that image to the Convolutional Neural Network (CNN) for processing. Features generated from that image are used for the classification of various environmental sound events such as sea waves, fire cracking, dog barking, lightning, raining, and many more. We have used the log-mel spectrogram auditory feature for training our six-layer stack CNN model. We evaluated the accuracy of our model for classifying the environmental sounds in three publicly available datasets and achieved an accuracy of 92.9% in the urbansound8k dataset, 91.7% accuracy in the ESC-10 dataset, and 65.8% accuracy in the ESC-50 dataset. These results show remarkable improvement in precise environmental sound recognition using only stack CNN compared to multiple previous works, and also show the efficiency of the log-mel spectrogram feature in sound recognition compared to Mel Frequency Cepstral Coefficients (MFCC), Wavelet Transformation, and raw waveform. We have also experimented with the newly published Rectified Adam (RAdam) as the optimizer. Our study also shows a comparative analysis between the Adaptive Learning Rate Optimizer (Adam) and RAdam optimizer used in training the model to correctly classifying the environmental sounds from image recognition architecture.

Index Terms: AESR, CNN, Log-Mel Spectrogram, MFCC, Adam, RAdam, Relu, Image, Classification.

1. Introduction

In recent times, automatic sound recognition has gained much impetus. It has been implemented in multidisciplinary areas like an audio surveillance system [1], detection of an impostor in the wildlife areas [2], environmental sound recognition in home automation [3], and extenuation of noise [4]. Various non-human sounds without music in our regular day to day life creates environmental sound, for example, glass breaking, door knock, pouring of water, the sound of an engine, helicopter, crying of a baby, and so many more. Recognition of such domestic and environmental sounds can open the door for innovative and significant business applications. The progression in the field of image classification generated from audio sound using deep learning techniques [5,6,7] are leading academicians and researchers to start using featured images created from the audio clip in a far more effective way to classify sound events. In AESR task, the objective is to recognize the type of a specific sound by labeling them into multiple events. These audio events recognition tasks are often assorted and cluttered with the acoustic scene classification [8] problem, where a sound feed to the neural network as input is required to be recognized into multiple acoustic scenes. However, we limit the range of our study to the recognition of sound events in an environment from audio clips.

Conventional audio features extraction methods such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) [9], Zero-Crossing Rate (ZCR), Wavelet Transformation have been used in the past to extract features from a sound event. However, the problem with MFCC is that it shows inadequacies in detecting sound when there is noise in the audio sample. LPC calculates in a linear structured way, so; it fails to take the non-linear features of an audio signal. ZCR is effectively used in detecting endpoint in musical appliances measurement. Nevertheless,

Measurement is functional only on longer fragments of the signal since small chunks might have just a few zero crossings or none at all. However, recent research suggests that Log-Mel Spectrogram (LMS) feature works better in detecting an environmental sound event. Last few years, Deep Neural Network (DNN) has made great success in automatic speech recognition and music information retrieval.

The Log-Mel Spectrogram (LMS) feature of an audio signal is considered as one of the most robust features for sound classification [10]. LMS is calculated for each frame of an auditory sample. Then a map is generated by forming the features of each frame along the time axis. CNN can be used to learn hidden patterns through a large amount of training data. Applying CNN to detecting environmental sound has gained an essential improvement by exploring log-mel spectrogram features.

Classification of sound from a featured image is one of the most trending topics in modern research. Orthodox machine learning algorithms such as Support Vector Machine (SVM), K Nearest Neighbor (KNN) & Gaussian Mixture Model (GMM) are already applied to detect and classify audio sound [11, 12].

There exists an exploration gap that concerns the competence of DNNs that was designed to identify objects when it comes to detecting a sound event based on spectrogram images. Our goal in this study is to improve the accuracy of detecting environmental sounds by proposing LMS based stack CNN model and hyper tuning the model with two types of padding in the CNN architecture and also make a comparative analysis of the model's performance by using Adam and Rectified Adam (RAdam) as the optimizer.

The remaining article is organized as follows. Section II describes the related work in AESR. In section III, we provide a brief depiction of the datasets used. Section IV presents the methodology and the network architecture of our study. In Section V, we describe the results of our experiments and the comparison of our results with other studies. Finally, we conclude our study and discuss real-world aspects of AESR in section VI.

2. Literature Review

With the advancement of Deep Convolutional Neural Network and its efficient use in computer vision, speech recognition, language modeling, and other related areas, it is proven that CNN based architecture outclasses the conventional methods in various classification tasks. Hence, they have been applied in the automatic sound event recognition task in recent years.

Piczak [13] evaluated the outcome of CNN based model in the AESR task by training on segmented spectrograms. He extracted log-mel features of each frame as an audio feature in a two-layer CNN model with max-pooling and two fully connected layers and achieved state of the art performance compared with the traditional methods.

Salamon & Bello [14] proposed a CNN model that contains three convolutional layers and one fully connected layer by using the log-mel feature as the two-channel input as well as augmenting data by pitch-shifting, time-stretching, dynamic range compression, and added background noise to the audio clips to increase the variety of data to train the model more efficiently. Their model improved the accuracy of classification by six percent on the urbansound8k dataset compared to Ref. [13].

CNN models like AlexNet and GoogleNet [15] initially developed for image recognition tasks has been exploited in many research works with state of the art outcome in environmental sound recognition.

Authors in Ref. [16], extracted the LMS, MFCC, and Cross Recurrence Plot (CRP) feature set from the sound clips and concatenated as a three-channel input to train the AlexNet and GoogleNet models with outstanding results in detecting environmental sound events.

In multiple research, raw waveforms were used to achieve the automatic learning of features in AESR [17, 18]. It is a combination of CNN that is used for extracting features from sound clips and Recurrent Neural Networks (RNN) for progressive aggregation of the features extracted using CNN. However, the results of the classification were unsatisfactory compared to the LMS feature.

Authors in Ref. [19], have developed a deeper classification network based on EnvNet, also referred to as EnvNet-v2, and achieved competitive performances using between class (BC) learning method. They generated the BC sounds by combining sounds from two diverse classes with a random ratio and trained their model with that.

Authors in Ref. [20], have proposed a network with dilated convolutional filters (enlarged filters) to gain more contextual information and more particular high-level features than traditional CNN architecture and experimented with their model with multiple activation functions to see which performed better in terms of classifying a specific sound event. According to their study, LeakyRelu achieved better accuracy than other activation functions with 81.9% accuracy in the urbansound8k dataset.

Uzkent, Barkana & Cevikalp in Ref. [21] introduced a new 2-D feature set based on Pitch Range (PR) of environmental sound, and an autocorrelation function used in the feature extracting method using Support Vector Machine (SVM). SVM classifier using the Gaussian kernel provided the highest accuracy of 85.6% in detecting non-speech environmental sound among the classifiers they have used.

Authors in Ref. [22] proposed a novel stacked CNN model which takes either raw waveform (RawNet) or log-mel (MelNet) feature as input. Using Dempster-Shafer (DS) evidence theory, they have developed an ensemble DS-CNN model by combining the previous two models for classifying environmental sound. Their MelNet model achieved 90.2% accuracy in the urbansound8k dataset, 91.4% accuracy in the ESC10 dataset, and 81.1% accuracy in the ESC50 dataset. Their RawNet model achieved 65.8% accuracy in the ESC50 dataset, 85.2% accuracy in the ESC10 dataset, and 87.7% accuracy in the urbansound8k dataset. Their Ave-CNN and Pro-CNN model achieved an accuracy of 91.6% and 91.9% in the urbansound8k dataset.

Authors in Ref. [23], have used log-mel featured based spectrogram images to train the CNN and Tensor Deep Stacking Network (TDSN) architecture. Their CNN model achieved an accuracy of 77% in the ESC10 dataset and 49% in the ESC50 dataset. Their TDSN model achieved an accuracy of 56% in the ESC10 dataset.

Authors in Ref. [24] proposed a novel robust optimization algorithm RAdam, which is a variant of the adaptive stochastic optimization algorithm, Adam. They have explored the warmup heuristic used for adaptive optimization algorithms. According to the authors, RAdam rectifies the adaptive learning rate of Adam to gain a more consistent variance.

Authors in Ref. [25] evaluated the performance of Machine Learning (ML) algorithms such as K-Nearest Neighbor (KNN), Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Tree (DT) to recognize urban sound on embedded devices concerning execution time and accuracy and they have also proposed a cascade approach to combine ML algorithms by analyzing the characteristics of embedded devices.

In this study, our objective is to progress the classification performance of the LMS based recognition of environmental sound by proposing a stacked CNN model and hyper tuning the model with two types of padding (same and valid) and also test the models' performance using two types of adaptive optimizer algorithm named Adam and Rectified Adam (RAdam).

3. Datasets

In our study, we have selected three publicly available datasets for evaluation of the model, ESC-50 [26], ESC-10 [26], and UrbanSound8K [27]. We wanted to train our model in a variety of sound samples of the environment from rural country place to nature, forest, sea, and urban society activities. Most of the previous researches conducted in this topic only worked with ESC-10 and ESC-50 datasets or only with the Urbansound8k dataset. We have selected all three in order to get more data for training the model with thousands of sound events of different categories. ESC-50 dataset contains two thousand short recorded audio clips of five seconds comprising of fifty equally sized classes. It is divided into major groups of animals, human non-speech sounds, interior/domestic sounds, natural soundscapes, water sounds, and urban noises and preset into five folds for cross-validation. The dataset consists of fifty .wav files sampled at 16 kHz for fifty diverse classes.

ESC-10 is a less complicated identical subclass with ten equally sized classes (dog barking, rain, sea waves, baby crying, a clock ticking, person sneezing, helicopter, chainsaw, rooster, fire crackling) of four hundred recordings selected from the ESC-50 dataset. Though the number of audio samples in the ESC-10 dataset is minimal to use for any deep learning techniques, we wanted to evaluate our model's performance in the real dataset without any augmentation.

UrbanSound8K is a pool of 8732 small sound clips (less than 4 seconds) of different urban sound (air conditioner, car horn, playing children, dog bark, drilling, engine idling, gunshot, jackhammer, siren, street music) preset into ten folds.

4. Methods & Architecture

A. Experimental Setup and Workflow of the study

We used a laptop with 16 GB RAM, an Intel Core i7-8750 CPU (8 cores @2.20 GHz), and NVIDIA GeForce GTX 1050-Ti graphics in the experiments. The PC was running Windows 10, and we have used Anaconda Python and the deep learning framework TensorFlow with Keras.

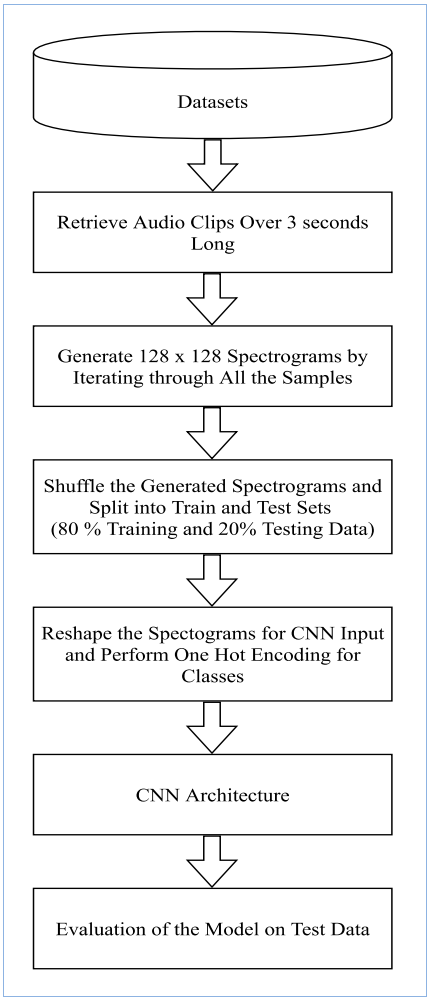

Fig.1. Basic workflow diagram of our study.

The basic workflow of our study is shown in figure 1. At first, we selected those audio clips, which were over three seconds long to generate high-level spectrogram featured images as valid data. We have used the librosa library in python to do that task for us. Then using librosa's melspectrogram feature, we generated 128x128 pixel log-mel spectrogram images by iterating through all the valid data and appended the class-ID of each data with its audio sample. It is done by dividing the frequency selection of audible sound into 128 components. It ranges from 0 to 22.05 kHz. There are 128 components across the time domain as well. It is divided into frames of 23.7 ms since each valid sound clips are over 3 seconds long. Hence, the overall input becomes a 128x128 matrix *A*, consisting of real numbers *R*, as shown in equation (1).

$$A_i \in R^{128 \times 128} \tag{1}$$

B. Network Architecture

The following tables 1 & 2 represents the architecture of the proposed CNN model for classifying environmental sound. Spectrogram image and wave plot representation of a sound event (siren) is provided in figure2.

Table 1. Summarization of the Proposed Architecture.

Input Shape	(128 x 128 x 1) log-mel spectrogram images.
Conv2D	16 kernels with a 3x3 receptive field.
Max Pooling	Size: (2x2)
Dropout	0.25
Activation function	Relu
Conv2D	32 kernels with a 3x3 receptive field.
Max Pooling	Size: (2x2)
Dropout	0.25
Activation function	Relu
Conv2D	64 kernels with a 3x3 receptive field.
Max Pooling	Size: (2x2)
Dropout	0.25
Activation function	Relu
Conv2D	128 kernels with a 3x3 receptive field.
Max Pooling	Size: (2x2)
Dropout	0.25
Activation function	Relu
Flatten(); Dropout: 0.5	
Fully Connected Layer	Dense(512)
Activation function	Relu
Dropout	0.6
Fully Connected Layer	Dense(10) for ESC-10 & Urbansound8K datasets and Dense(50) for ESC-50 dataset.
Activation function	Softmax

We evaluated the model with K-fold cross-validation in all the datasets with K=5 for ESC-50 and ESC-10 datasets and K=10 for the urbansound8k dataset. We have used dropout in each layer to prevent overfitting.

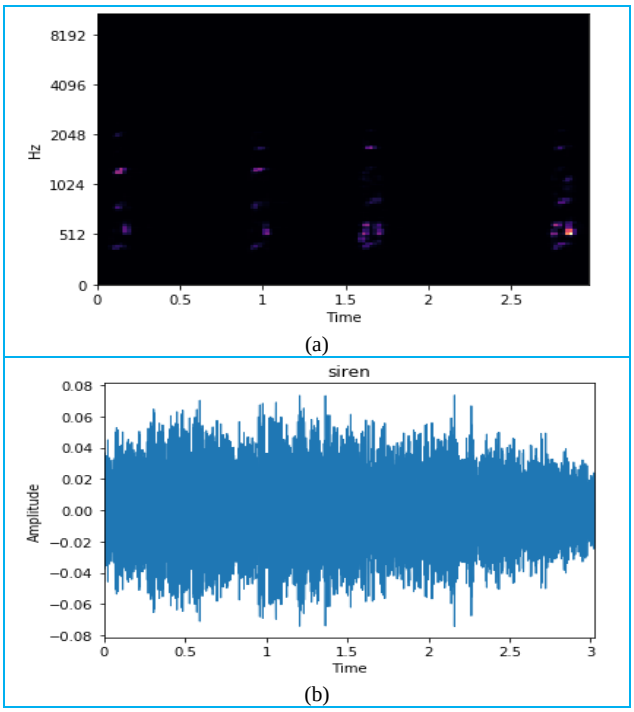

Fig.2. (a) Spectrogram image of a siren, (b) wave plot of a siren.

Table 2. The architecture shape and number of parameters of the proposed six layers stack CNN model.

Layer (Type)	Output Shape	Number of Parameters
conv2d_1 (Conv2D)	(None, 126, 126, 16)	160
max_pooling2d_1 (MaxPooling2D)	(None, 63, 63, 16)	0
dropout_1 (Dropout)	(None, 63, 63, 16)	0
activation_1 (Activation)	(None, 63, 63, 16)	0
conv2d_2 (Conv2D)	(None, 63, 63, 32)	4640
max_pooling2d_2 (MaxPooling2D)	(None, 31, 31, 32)	0
dropout_2 (Dropout)	(None, 31, 31, 32)	0
activation_2 (Activation)	(None, 31, 31, 32)	0
conv2d_3 (Conv2D)	(None, 31, 31, 64)	18496
max_pooling2d_3 (MaxPooling2D)	(None, 15, 15, 64)	0
dropout_3 (Dropout)	(None, 15, 15, 64)	0
activation_3 (Activation)	(None, 15, 15, 64)	0
conv2d_4 (Conv2D)	(None, 15, 15, 128)	73856
max_pooling2d_4 (MaxPooling2D)	(None, 7, 7, 128)	0
dropout_4 (Dropout)	(None, 7, 7, 128)	0
activation_4 (Activation)	(None, 7, 7, 128)	0
flatten_1 (Flatten)	(None, 6272)	0
dropout_5 (Dropout)	(None, 6272)	0
dense_1 (Dense)	(None, 512)	3211776
activation_5 (Activation)	(None, 512)	0
dropout_6 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 10)	5130
activation_6 (Activation)	(None, 10)	0
Total params: 3,314,058 Trainable params: 3,314,058 Non-trainable params: 0		

5. Result Analysis

A. Model's performance in the datasets:

Two types of padding (same and valid) were used in the convolutional layer. We tested the model on both Adam and RAdam optimizer. In our experiment in recognizing environmental sound from log-scaled mel spectrogram featured images, Adam optimizer performed better than Rectified Adam (RAdam) optimizer in terms of correctly classifying sound events. The batch size was set at 32 when fitting the model. In every training of our model, we noticed that though RAdam showed robust characteristics in its heuristics to obtain a more stable variance, its accuracy in detecting sound was lesser than Adam optimizer in every training.

Table 3. Performance chart of the model in the Urbansound8k dataset.

Padding	Optimizer	Accuracy	No. of epochs
Same	Adam	92.9%	220
Valid	Adam	89%	
Same	RAdam	82.5%	
Valid	RAdam	77%	

Table 4. Performance chart of the model in the ESC-10 dataset.

Padding	Optimizer	Accuracy	No. of epochs
Same	Adam	91.7%	1000
Valid	Adam	81%	
Same	RAdam	82.2%	
Valid	RAdam	75%	

Table 5. Performance chart of the model in the ESC-50 dataset.

Padding	Optimizer	Accuracy	No. of epochs
Same	Adam	65.8%	1000
Valid	Adam	57%	
Same	RAdam	44%	
Valid	RAdam	33%	

From our study, we find that the accuracy of the model decreased significantly when valid padding was used in the architecture. The validation accuracy curve in each training of the experiment with the highest accuracy in both optimizers shown in tables 3-5 is presented in figure no. 3 to figure no. 8.

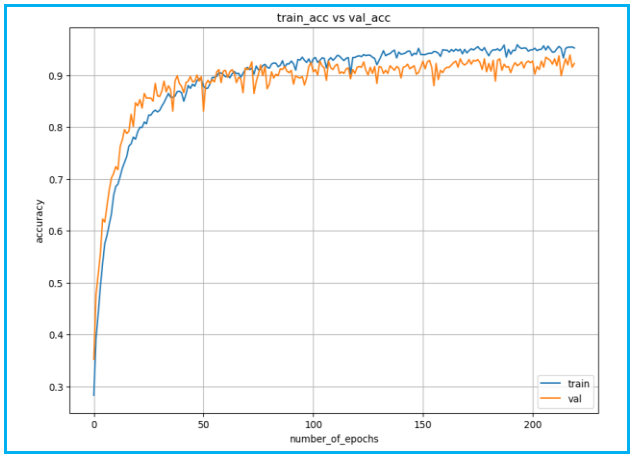

Fig.3. Training accuracy vs. Validation accuracy curve of the model in the urbansound8k dataset with Adam optimizer and the same padding in the convolution layer with a validation accuracy of 92.9%.

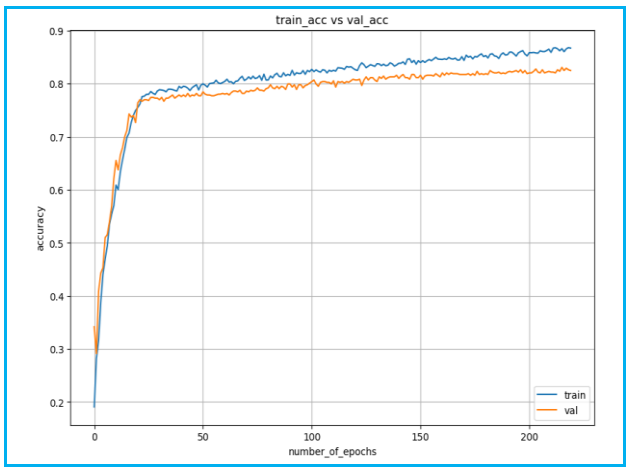

Fig.4. Training accuracy vs. Validation accuracy curve of the model in the urbansound8k dataset with Rectified Adam optimizer and same padding in the convolution layer with a validation accuracy of 82.5%.

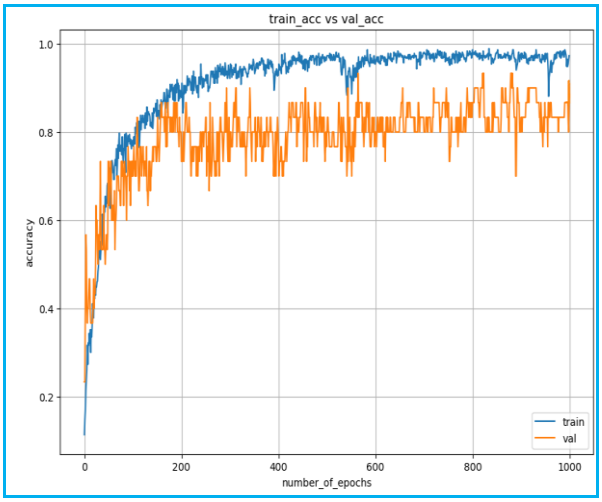

Fig.5. Training accuracy vs. Validation accuracy curve of the model in the ESC-10 dataset with Adam optimizer and the same padding in the convolution layer with a validation accuracy of 91.7%.

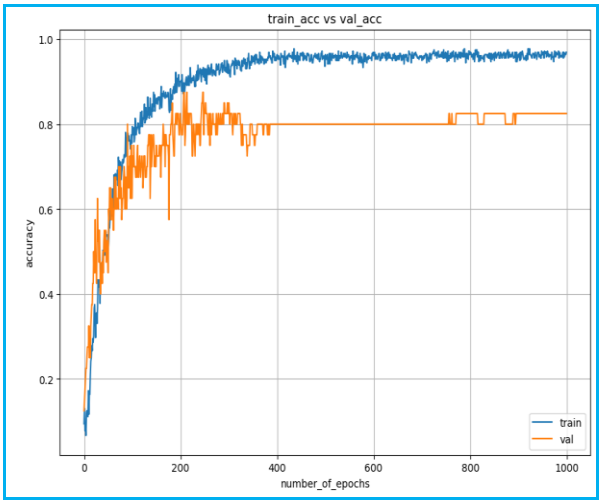

Fig.6. Training accuracy vs. Validation accuracy curve of the model in the ESC-10 dataset with Rectified Adam optimizer and same padding in the convolution layer with a validation accuracy of 82.2%.

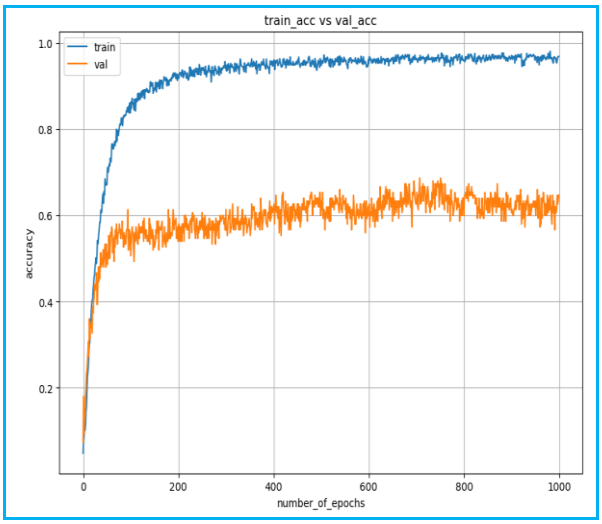

Fig.7. Training accuracy vs. Validation accuracy curve of the model in the ESC-50 dataset with Adam optimizer and the same padding in the convolution layer with a validation accuracy of 65.8%.

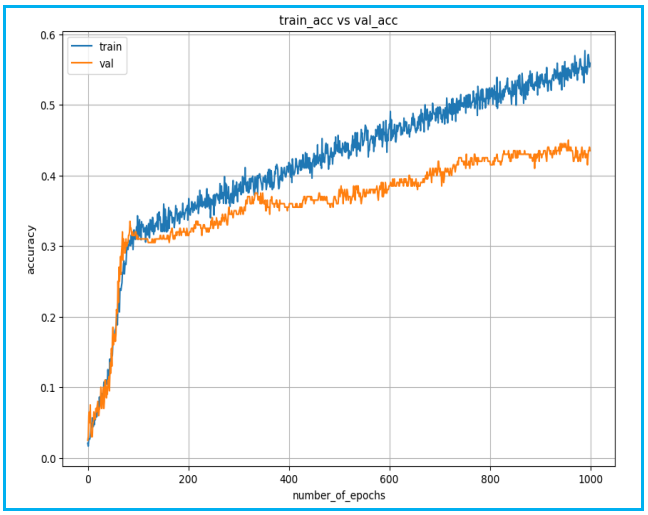

Fig.8. Training accuracy vs. Validation accuracy curve of the model in the ESC-50 dataset with Rectified Adam optimizer and same padding in the convolution layer with a validation accuracy of 44%.

The normalized confusion matrixes of the model in Urbansound8k, ESC-50, and ESC-10 datasets, respectively, are presented in figures 9 to 11. The confusion matrix is a summary of the prediction results on a classification problem. Diagonal elements of each matrix show the validation accuracy of predicting the environmental sound of specific sound events or classes.

Table 6. Functional chart of a confusion matrix

		Predicted Data	
		(Positive: P)	(Negative: N)
Actual Data	(Positive: P)	TP	FN
	(Negative: N)	FP	TN

We can calculate the accuracy of the proposed CNN model as the number of all accurate predictions divided by the total number of valid data presented the dataset using equation (2).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

Here, TP = True Positive event prediction, TN = True Negative event prediction, FP = False Positive event prediction, and FN = False Negative event prediction.

Validation Accuracy on Urbansound8k dataset from figure 9:

$$Accuracy = \frac{(.87 + .95 + .91 + 1 + .94 + 1 + .89 + .86 + .9 + .97)}{10} * 100\% \quad Accuracy = 92.9\%$$

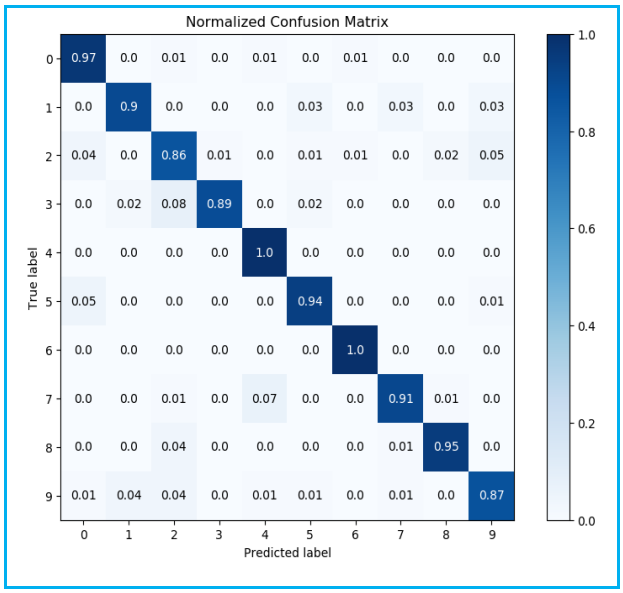

Fig.9. Class-wise normalized confusion matrix of the model in the **urbansound8k** dataset with Adam optimizer and padding (same) in the convolution layer with a validation accuracy of **92.9%**. Here, labels, 0=Air conditioner, 1=Car horn, 2= Children playing, 3=Dog bark, 4=Drilling, 5=Engine idling, 6=Gun shot, 7=Jackhammer, 8=Siren, and 9=Street music [dataset: Urbansoound8k].

Validation accuracy in ESC-10 dataset from figure 10,

$$Accuracy = \frac{(1+1+1+0.67+1+0.5+1+1+1+1)}{10} * 100\%$$

$$Accuracy = 91.7\%$$

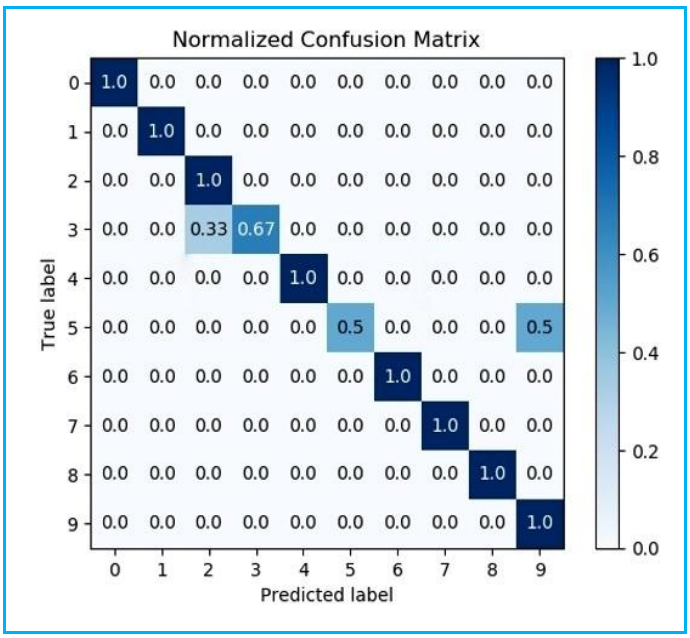

Fig.10. Class-wise normalized confusion matrix of the model in the **ESC-10** dataset with Adam optimizer and padding (same) in the convolution layer with a validation accuracy of **91.7%.** Here, labels, 0=dog, 1=rooster, 2=rain, 3=sea waves, 4=crackling fire, 5=crying baby, 6=sneezing, 7=clock tick, 8=helicopter, and 9=chainsaw [dataset: ESC10]. The model correctly predicts the sound of dog, rooster, rain, crackling fire, sneezing, clock tick, the helicopter, and chainsaw with 100% accuracy.

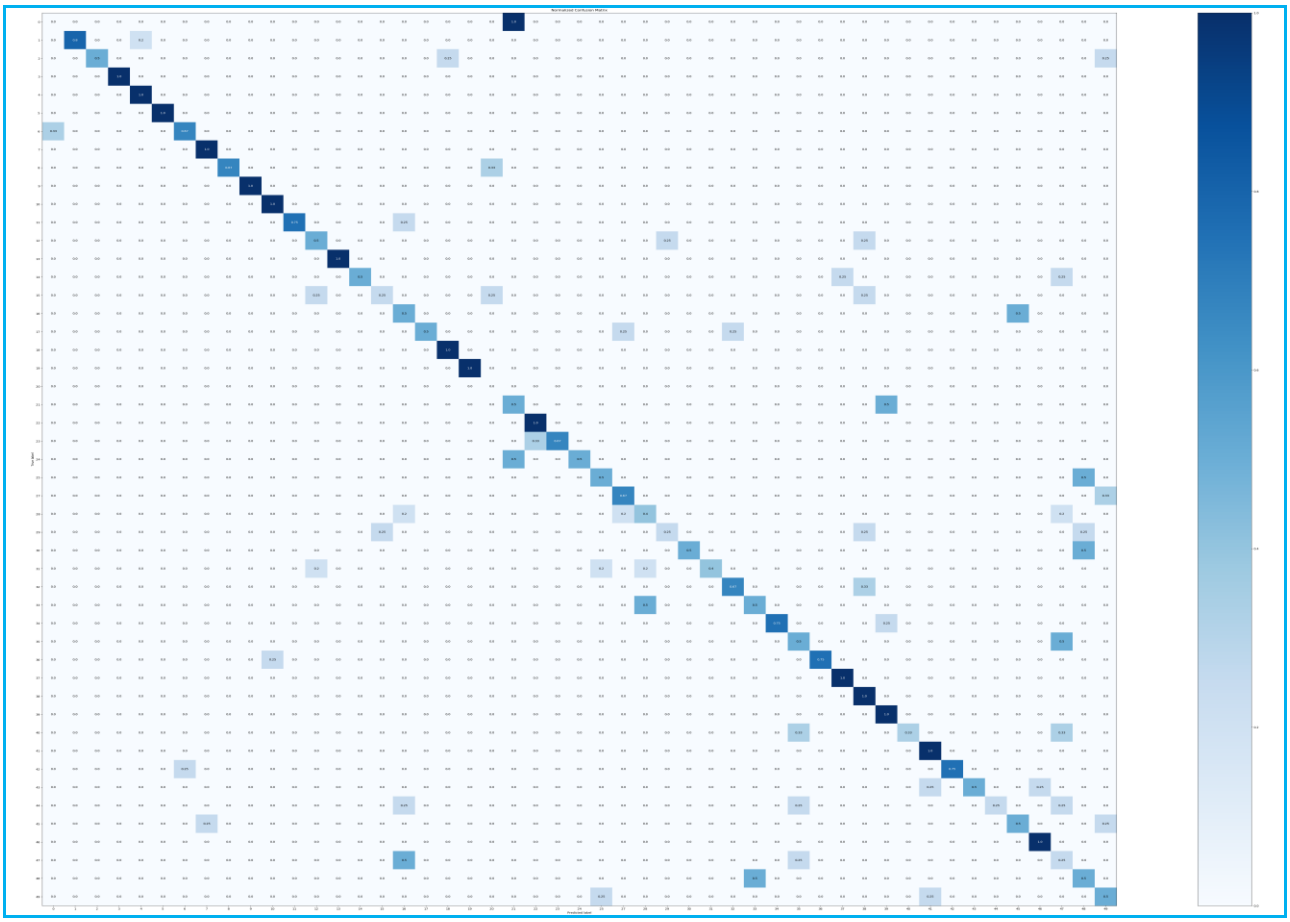

Fig.11.Class-wise normalized confusion matrix of the model in the ESC-50 dataset with Adam optimizer and padding (same) in the convolution layer with a validation accuracy of **65.8%** [dataset: **ESC-50**]. The model correctly classifies label 0=cow, 4=frog, 5=cat, 7=insects, 9=crow, 10=rain, 13=crickets, 18=toilet flush, 19=thunderstorm, 22=clapping, 37=clock alarm, 38=clock tick, 39=glass breaking, 41=chainsaw and 46=church bells sound with 100% accuracy. It classifies label 1=rooster, 34=can opening, 36= vacuum cleaner, and 11=sea waves sound between 67% to 80% accuracy. The classification accuracy of the rest of the classes was between 25% to 50%. The model somehow failed to predict the sound of crying baby and dog and gets confused with water drop, sheep, and sneezing in the ESC-50 dataset.

B. Comparison of our model with other baseline models

In tables 7, 8, 9, we have compared our result with other states of the art neural network models. In the Urbansound 8k dataset, our approach yields the best accuracy of 92.9%. In the ESC 10 and ESC 50 dataset, our accuracy is 91.7% and 65.8%, respectively.

Table 7. Comparison of classification accuracy with other baseline models on the Urbansound8k dataset. Bold mark with green background indicates our result.

Model	Feature	Accuracy
Piczak [13]	Log-mel	73.7%
Salamon [14]	Log-mel	79%
AlexNet [15]	Log-mel	92%
Boddapati [16]	MFCC+CRP+Spectrogram	93%
Dai [18]	Raw waveform	71.68%
Zhang [20]	Log-mel	81.9%
MelNet [22]	Log-mel	90.2%
RawNet [22]	Raw waveform	87.7%
DS-CNN [22]	Combine (DS evidence)	92.2%
Ave-CNN [22]	Combine (Average)	91.6%
Pro-CNN [22]	Combine (Product of probabilities)	91.9%
Our Approach	**Log-mel**	**92.9%**

Table 8. Comparison of classification accuracy with other baseline models on the ESC-10 dataset. Bold mark with green background indicates our result.

Model	Feature	Accuracy
Piczak [13]	Log-mel	81.5%
RawNet [22]	Raw waveform	85.2%
Khamparia [23]	Log-mel	77%
Our Approach	**Log-mel**	**91.7%**

Table 9. Comparison of classification accuracy with other baseline models on the ESC-50 dataset. Bold mark with green background indicates our result

Model	Feature	Accuracy
EnvNet [10]	Raw waveform	64%
Piczak [13]	Log-mel	64.5%
RawNet [22]	Raw waveform	65.7%
Khamparia [23]	Log-mel	53%
Our Approach	**Log-mel**	**65.8%**

6. Conclusion

In our paper, we have projected an intelligent stack CNN model to recognize environmental sound events that is one of the less explored areas in the present research field. We used the Log-Mel (LM) spectrogram as the primary feature and generated LM-spectrogram images for each valid sound clip in the evaluated datasets. We have used three public datasets, Urbansound8k, ESC-50, and ESC-10, to assess the classification performance of the model. We performed multiple hyper tuning of the model with different dropout rates, different types of padding in the convolution layer, changing the max-pooling layer size as well as tuning with multiple stride steps to see which combination provides the best accuracy in recognizing environmental sound efficiently. We evaluated the model with Adam optimizer and Rectified Adam optimizer. Though Rectified Adam optimizer has a better warmup heuristic to achieve the highest accuracy in a shorter number of epochs standard Adam optimizer performed better in terms of providing better classification accuracy. Our model outperformed multiple baseline models in the automatic environmental sound reorganization task with an accuracy of 92.9% in the Urbansound8k, 91.7% in the ESC-10, and 65.8% in the ESC-50 datasets. However, trainable parameters of 3.3M provide a greater computation complexity and more significant memory. In the future, we will put our sight to increase the classification accuracy of the model, even more, to recognize more detailed sound samples with varying frequency and signal to noise ratio as well as dropping the computational cost of the model. This system can be applied as a hearing aid for different environment settings and could open the door for multiple practical applications used on embedded systems for commercial purposes.

Acknowledgment

The authors would like to acknowledge the anonymous reviewers for their valued recommendations. The authors did not receive any kind of grant for this study to be carried out.

References

[1] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, “Using one-class SVMs and wavelets for audio surveillance,” IEEE Transactions on information forensics and security 3 (4), 763–775, 2008.
[2] M. V. Ghiurcau, C. Rusu, R. C., Bilcu, J. Astola, “Audio based solutions for detecting intruders in wild areas,” Signal Processing 92 (3), 829–840, 2012.
[3] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, “Robust Environmental Sound Recognition for Home Automation,” IEEE Transactions on Automation Science and Engineering, vol. 5, no. 1, pp. 25–31, Jan. 2008.
[4] Mydlarz, C.; Salamon, J.; Bello, J.P. “The implementation of low-cost urban acoustic monitoring devices,” in Appl. Acoust. 2016, 117, 207–218.
[5] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 559-563.
[6] O. Gencoglu, T. Virtanen and H. Huttunen, “Recognition of acoustic events using deep neural networks,” in 2014 22nd European Signal Processing Conference (EUSIPCO), Lisbon, 2014, pp. 506-510.
[7] S. Chachada and C. -. J. Kuo, “Environmental sound recognition: A survey,” in 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kaohsiung, 2013, pp. 1-9.

[8] K. Yao, J. Yang, X. Zhang, C. Zheng and X. Zeng, "Robust Deep Feature Extraction Method for Acoustic Scene Classification," in 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 2019, pp. 198-202.
[9] E. R. Swedia, A. B. Mutiara, M. Subali, and Ernastuti, "Deep Learning Long-Short Term Memory (LSTM) for Indonesian Speech Digit Recognition using LPC and MFCC Feature," in 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-5.
[10] Tokozume, Yuji, and T. Harada. "Learning environmental sounds with end-to-end convolutional neural network," 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2721-2725. IEEE, 2017.
[11] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," in IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16-34, May 2015.
[12] Theodorou, T. Mporas, I. Fakotakis, N, "Automatic Sound Recognition of Urban Environment Events," Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 129–136
[13] K. J. Piczak, "Environmental sound classification with convolutional neural networks," 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, 2015, pp. 1-6.
[14] Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017.
[15] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.
[16] Boddapati, Venkatesh, Andrej Petef, Jim Rasmusson, and Lars Lundberg, "Classifying environmental sounds using image recognition networks," Procedia computer science 112 (2017): 2048-2056.
[17] J. Sang, S. Park and J. Lee, "Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms," 2018 26th European Signal Processing Conference (EUSIPCO), Rome, 2018, pp. 2444-2448.
[18] Dai, W. Dai, C. Qu, S. Li, J. Das, S. "Very deep convolutional neural networks for raw waveforms," Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 421–425.
[19] Tokozume, Y., Ushiku, Y., Harada, T. "Learning from between-class examples for deep sound recognition," arXiv preprint arXiv: 1711.10282 (2018).
[20] X. Zhang, Y. Zou and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," 2017 22nd International Conference on Digital Signal Processing (DSP), London, 2017, pp. 1-5.
[21] Uzkent, Burak, Buket D. Barkana, and Hakan Cevikalp, "Non-speech environmental sound classification using SVMs with a new set of features," International Journal of Innovative Computing, Information and Control 8, no. 5 (2012): 3511-3524.
[22] Li, Shaobo, Yong Yao, Jie Hu, Guokai Liu, Xuemei Yao, and Jianjun Hu. "An ensemble stacked convolutional neural network model for environmental event sound recognition." Applied Sciences 8, no. 7 (2018): 1152.
[23] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey and P. Tiwari, "Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," IEEE Access, vol. 7, pp. 7717-7727, 2019.
[24] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, "On The Variance of the adaptive learning rate and beyond," arXiv:1908.03265v2 [cs.LG] 10 Mar 2020.
[25] da Silva, Bruno, Axel W Happi, An Braeken, and Abdellah Touhafi. "Evaluation of Classical Machine Learning Techniques towards Urban Sound Recognition on Embedded Systems." Applied Sciences 9, no. 18 (2019): 3885.
[26] Piczak, Karol J. "ESC: Dataset for environmental sound classification." Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018. 2015.
[27] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041-1044. 2014.

Authors' Profiles

Md. Rayhan Ahmed is currently serving as a Senior Lecturer in the Department of Computer Science and Engineering at Stamford University Bangladesh. He received his Bachelor of Science degree from Ahsanullah University of Science and Technology (AUST) in 2014. He is pursuing his Master of Science degree at United International University (UIU), Bangladesh. His research interest is Machine Learning, Artificial Intelligence, Deep Learning, Computer Vision, Data Science, Implementation of the Internet of Things (IoT) in real-world applications, and development of the android application.

Md. Towhidul Islam Robin is currently serving as a Senior Lecturer in the Department of Computer Science and Engineering at Stamford University Bangladesh (SUB). He received his Bachelor of Science degree from Ahsanullah University of Science and Technology (AUST) in 2015. He is currently pursuing a Master of Science Degree in Computer Science and Engineering (CSE) from United International University (UIU), Dhaka, Bangladesh. His research interest lies within Machine Learning, Data Mining, Natural Language Processing (NLP), and the Internet of Things.

Ashfaq Ali Shafin is appointed as a Lecturer for the Department of Computer Science and Engineering at Stamford University Bangladesh. In 2018, he earned his Bachelor of Science degree in Computer Science and Engineering from Ahsanullah University of Science and Technology (AUST), securing the second position in a class of 112 students. At present, he is exploring the research fields of Machine Learning, Natural Language Processing (NLP), Digital Image processing (DIP), and Computer Vision (CV).

How to cite this paper: Md. Rayhan Ahmed, Towhidul Islam Robin, Ashfaq Ali Shafin, " Automatic Environmental Sound Recognition (AESR) Using Convolutional Neural Network", International Journal of Modern Education and Computer Science(IJMECS), Vol.12, No.5, pp. 41-54, 2020.DOI: 10.5815/ijmecs.2020.05.04

