Sound Event Recognition and Classification Using Machine Learning Techniques

SBK Karunaratna# and MWP Maduranga

Department of Computer Engineering, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka

#D-CE-18-0004@kdu.ac.lk

Abstract **–** Sound event recognition and classification are exciting and vital applications in the era of the Internet of Things (IoT). These Sound events carry information that is useful for our daily lives. The perception of surrounding events by humans depends strongly on audio signals. Awareness of what happens in the surrounding environment depends heavily on the ability of an individual to perceive sounds and accurately recognize events related to them. The subject of audio signal recognition is now very popular and has numerous applications. This paper presents machine learning approaches to classify sound events extracted through sound sensors, where the sound signals acquired by sensors will be processed using machine learning algorithms to classify them. The results show that the accuracy of CNN, SVM, MLP classifiers are 82%, 81%, and 79.48%, respectively.

Keywords: sound event recognition, Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Multilayer Perception (MLP)

I. INTRODUCTION

The world is filled with sounds that are carrying information. Sounds are essential for the execution of regular activities, social interaction, and even personal protection. It provides information about character, place, and time. Sounds can inform and move us in ways that visuals alone cannot. The topic of audio signal recognition has received a lot of attention in recent years. Audio signals can be split into three categories: music, voice, and sound. Out of them, sound event signal detection has recently become more critical in the general daily environment. Sound event recognition has many applications. It can be used for elderly care, to build devices for people with hearing loss, for defense and surveillance purposes, for identifying sounds in emergency and notifying authorized people, such as the house's fire alarm or someone crying for help, and many more. Machine learning approaches can be used to give a human-like performance on sound event recognition and classification. This paper presents our investigation on popular machine learning techniques for sound event recognition and classification.

Sound event recognition consists of three stages: sound acquisition, feature extraction, and classification. First, most audio signal identification systems nowadays use a microphone array or a camera to receive sound signals. A large number of high-sensitivity microphones are required for microphone arrays. Hence microphone arrays are expensive. Furthermore, it is time-consuming to analyze the signals received by the microphone array. A camera is also costly, and interference of objects can happen. Therefore, we have used only a single microphone to receive the sound signals and detect the sound events in this paper. Second, the features of signals were extracted using MFCC (Mel-frequency cepstrum). Finally, different Machine learning techniques were used for classification.

In this work, the ESC-50 dataset was used for model training [1]. It consists of 50 different sound events. Out of them, ten sound events were defined: dog barking, baby crying, siren, glass breaking, fire crackling, rain, washing machine, car horn, clock alarm, and birds chirping.

II. RELATED WORK

Mete Yaganoglu and camal kose proposed a wearable device for hard-of-hearing people detecting sounds and notifying the user using vibration patterns. The use of different vibration

patterns made the system complicated and hard to understand. It uses a KNN algorithm for classification.[2]

Tariq, Z., Shah, S., and Lee, Y. e discussed speech emotion analysis and how emotion can be used in medicine, a psychological study on old adults in nursing homes. Further discussed is an IoT-based speech emotion detection system that can process real-time audio to predict the emotion of aging adults. A 2D CNN model was used together with RMS, Peak, and EBU normalization and data augmentation techniques [3].

Wang J, Lee, S., Fu, Z., and Shih, P. suggested a method for Robust Sound Recognition Applied to Awareness for Health/Children/Elderly Care. This research uses an ICA-transformed MFCCs feature frame-based multi-class SVM. An accuracy of 90.97 percent was achieved [4].

This paper presents our investigations on automatic daily sound recognition using ensemble methods. First individual classifiers are used with different acoustic features to measure their performance in recognizing everyday sounds. Ensemble methods are then employed to identify the daily sounds better [5].

Chang, C., and Chang performed research on recognizing abnormal indoor sounds; Y. SVM algorithms were used as the classifier. Four discriminative features (peak, valley, and contrast of Octave-based and MFCC) were chosen. An accuracy of 86 percent was achieved [6].

III. DATASET

The dataset was prepared using recordings of the ESC-50 dataset for model training. The ESC-50 dataset consists of 5 - second long 2000 environment recordings of 50 different sound events. Out of them, ten various sound events were selected. They are baby crying, siren, glass breaking, rain, washing machine, car horn, clock alarm, birds chirping, fire crackling, dog bark. The dataset consists of 5 second long 400 recordings. Each class has 40 recordings. The data set is divided into 80% training data and 20% testing data**F\\\EATURE EXTRACTION**

When it comes to classification, features are crucial. They represent the sound in a simplified numerical format. To make the classification process more effective and accurate, we have to extract features from the audio clips. MFCC (Mel-frequency cepstrum) is a standard feature extraction method for sound recognition because of its high efficiency. The basic procedure to develop MFCCs is,

i) Convert from Herts to Mel Scale

$$m = 1127 \, x \log\left(1 + \frac{f}{700}\right) \quad (1)$$

ii) Take the logarithm of the Mel representation of the audio

$$\log E = \log\left(\text{sum}(x^2)\right) \text{ (Log of Time-domain)} \quad (2)$$

iii) Take the logarithmic magnitude and use Discrete Cosine Transformation

$$F(u) = \left(\frac{2}{N}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} A(i) \cos\left[\frac{\pi . u}{2.N}(2i+1)\right] f(i) \quad (3)$$

This creates a spectrum over Mel frequencies as opposed to time, thus creating MFCCs.

Here we use librosa's mfcc() function, which generates an MFCC from time series audio data to extract the MFCC for each audio file (Ogg) in the dataset. [7]

The sound files are read using the Soundfile library. Then MFCC is extracted from each sound

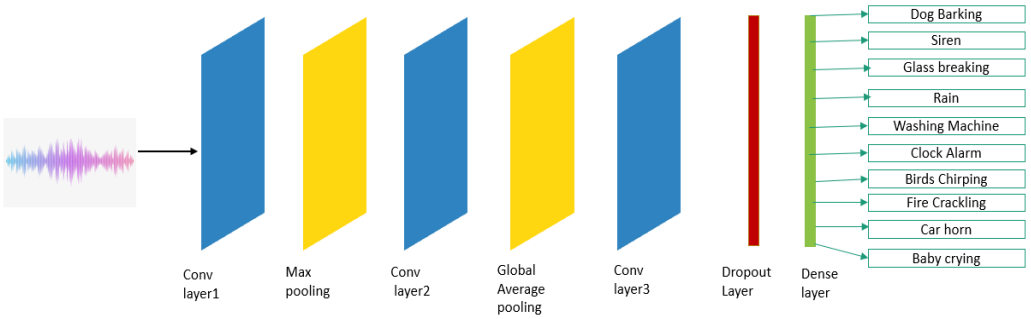

Figure 8: CNN Model

file. Finally, a data frame is created with the MFCC feature and corresponding class label. Once the feature extraction is done, it is converted into a NumPy array and taken as input to the classifiers.

IV. RECOGNITION OF SOUND

A study was done to find which ML algorithms are most suitable for sound event recognition and classification. It was found that mainly convolutional neural networks (CNN), Support vector machine (SVM), Multilayer perception (MLP) algorithms are used for Sound event recognition and classification. So, the data set was trained and tested on these classification models, and the accuracies were taken.

A. CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network is gaining popularity in deep learning, and its use in the audio domain is increasing. CNN is a well-known model in the fields of image, text, and other computer vision applications. We have implemented a 1D CNN model using the Keras framework. A convolutional neural network has two main components: a feature extractor and a classifier. The feature extractor extracts the MFCC from the audio signal and sends them to a classifier for classification. The classifier is made up of various convolutional and pooling layers, which are then followed by activation. Following Activation, Functions are used in this project.

ReLU Activation Function –

$$f(x) = \max(0, x) \quad (4)$$

Softmax Activation Function –

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (5)$$

The 1D CNN architecture is composed of 8 layers. The first layer is made up of a 40x1 input image. Then it is convolved with 64 filters of size three kernel in the next layer. Then the second layer is a Max Pooling Operation. Then in the third layer, it is again convolved with 128 filters of size 3. All of the convolution layers are followed by ReLU as the activation function. Then the fourth layer is a Global Average Pooling Operation. Next, a dropout layer of rate 0.5 was used to stop overfitting. The last layer is a Softmax output layer with the ten classes in the dataset. The model is compiled using Adam optimizer. The Total accuracy obtained after 100 epochs with a batch size of 32 is 84%.

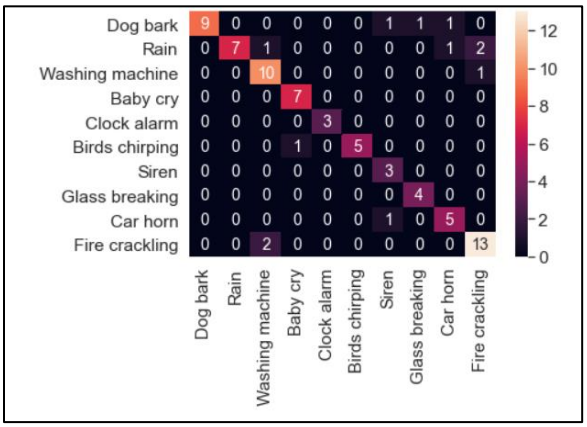

Figure 10 confusion matrix of CNN test data

B. SVM CLASSIFIER

A robust and popular classification machine nowadays is the support vector machine (SVM). Its idea is simple; the algorithm builds a dividing line to divide data into classes. SVM takes data as an input and produces a hyperplane which, if possible, separates these classes. SVM attempts to make a decision boundary so that there is a maximum distance between the two categories. In SVM models, several kernels can be used. These include polynomial, RBF, and Sigmoid function. In this project, the kernel type was taken as a second-degree polynomial. Thus, a higher-order polynomial sets out the decision boundary that separates classes.

Below is the polynomial function.

$$K(X_1, X_2) = (a + X_1^T X_2)^b \quad (6)$$

where b is the degree of the polynomial

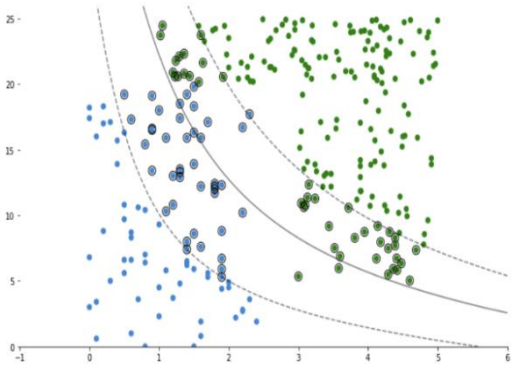

Figure 9: A conceptual diagram of a second-degree polynomial

A grid search algorithm is applied to find the best parameters of C to improve the classification accuracy of the SVM. The total accuracy of the SVM classifier is 81%.

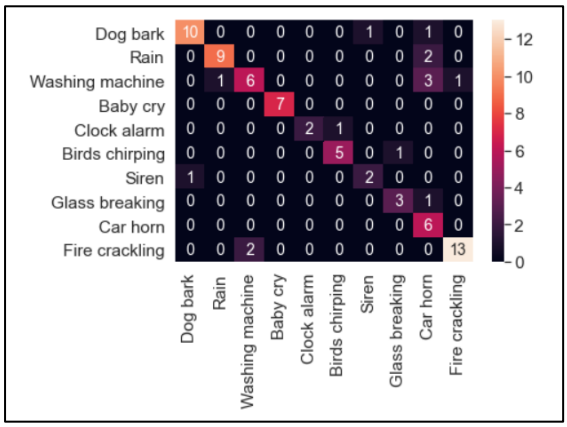

Figure 11 Confusion matrix of SVM test data

C. *MULTILAYER PERCEPTRON*

Multilayer perception (MLP) is a class of feedforward artificial neural networks (ANN). An MLP has at least three nodes: one input layer, one covered layer, and one output layer. MLP uses a supervised learning method known as backpropagation for training. The below figure illustrates a simple, fully connected two-layer MLP network.

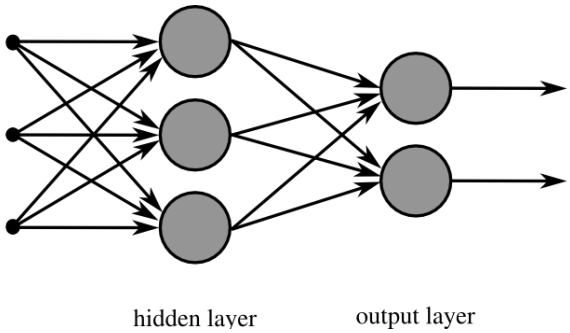

Figure 12 A conceptual diagram of an MLP

The mlp created for this project consists of 3 fully connected layers. The layers have 256, 256, and 10 neurons in them. Relu activation function is used in all layers except the last layer. The softmax activation function is used in the previous layer. In the first two layers, dropout layers have been used to reduce the overfitting of the training data. The model is compiled using Adam optimizer. Keras was used to build the MLP. The MLP was trained for 200 epochs with a batch size of 32. The total accuracy of the MLP classifier is 79.48%

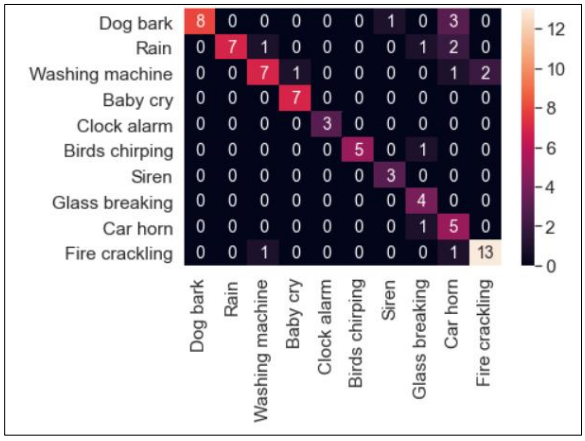

Figure 13:Confusion matrix of mlp test data

V. RESULTS AND ANALYSIS

We trained and tested CNN, SVM, and MLP learning algorithms to identify the sound events accurately in this experiment. For the experiment, we used **ESC-50 data set, which consists of 50 sound events and among them 10 number events, namely "baby crying," "dog barking,"** "siren," "glass breaking," "fire crackling," "rain," "washing machine," "car horn," "clock alarm" and "birds chirping**," were tested. During the testing phase, 80% were selected as the training set, and 20% were used as the test set. Data pre-processing were done using** MFCC (Mel-frequency cepstrum), and features were extracted. **The results show that** the total accuracies of the CNN, SVM, MLP classifiers are 84%, 81%, and 79.48%, respectively. Out of the three models, the CNN model has the highest accuracy of 82%. So, the convolutional neural network has outperformed the other two models.

VI. CONCLUSIONS

In this paper, we discussed how popular machine learning techniques work for the classification of sound events. The **ESC-50 data set was used for the training and testing. The experiment was carried out identifying sound events "baby crying," "dog barking,"** "siren," "glass breaking," "fire crackling," "rain," "washing machine," "car horn," "clock alarm" and "birds chirping**," with the** approaches for sound event recognition and classification. MFCC (Mel-frequency cepstrum) features were extracted from the signals. Three machine learning algorithms were discussed in this paper. They

are SVM, CNN, and MLP. The CNN architecture consists of 8 layers followed by ReLU and Softmax activation functions. It was trained for 100 epochs with a batch size of 32. The SVM uses a second-degree polynomial kernel. A grid search algorithm is applied to find the best parameters of C to improve the classification accuracy of the SVM. The MLP architecture consists of three fully connected layers. ReLU and Softmax activation functions also follow these layers. The MLP was trained for 200 epochs with a batch size of 32. Both the CNN and MLP models use the Adam optimizer for compilation. From the models, CNN, SVM, MLP, the following accuracies were obtained as 84%, 81%, and 79%, respectively. The CNN model has the highest accuracy compared to the other two models, so it outperformed in terms of accuracy.

REFERENCES

K. J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In: *Proceedings of the 23rd Annual ACM Conference on Multimedia, Brisbane, Australia.*

Yağanoğlu, M. and Köse, C., 2018. Real-Time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People. *Electronics*, 7(4), p.50.

Tariq, Z., Shah, S. and Lee, Y., 2019. Speech Emotion Detection using IoT based Deep Learning for Health Care. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE.

Wang, J., Lee, S., Fu, Z. and Shih, P., 2011. Robust Sound Recognition Applied to Awareness for Health/Children/Elderly Care. In: *2011 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE.

Shaukat, A., Ahsan, M., Hassan, A. and Riaz, F., 2014. Daily Sound Recognition for Elderly People Using Ensemble Methods. In: *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*.

Chang, C. and Chang, Y., 2013. Application of Abnormal Sound Recognition System for Indoor Environment. In: *2013 9th International Conference on Information, Communications & Signal Processing*. IEEE.

Iwaki, M. and Nakayama, S., 2018. Sound-Recognition Method for Helping Us Respond Appropriately to Sounds in Daily Life. In: *2018 IEEE 7th Global Conference on Consumer Electronics*. IEEE.

McFee, B. et al., 2015. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference.

Latif, G., Khan, A. and Butt, M., 2017. IoT based Real-time Voice Analysis and Smart Monitoring System for Disabled People. In: *1st International Conference on Advanced Research (ICAR -2017)*.

McFee, B., Raffel, C., Liang, D., Ellis, D., Nieto, O. and McVicar, M., 2015. librosa: Audio and Music Signal Analysis in Python. In: *14th PYTHON IN SCIENCE CONF. (SCIPY 2015)*.

Salamon J. and Bello j. p.,2017. Deep convolutional neural networks and data augmentation for environmental sound classification, In 2017 IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283. IEEE

Bempong, J., Stainslow, J. and Behm, G., 2015. Accessible Smart Home System for the Deaf and Hard-of-Hearing.

Mushtaq, Z. and Su, S., 2020. Efficient Classification of Environmental Sounds through Multiple Features Aggregation and Data Enhancement Techniques for Spectrogram Images. *Symmetry*, 12(11), p.1822.

Yoo, I. and Yook, D., 2008. Automatic Sound Recognition for the Hearing Impaired. In: *IEEE Transactions on Consumer Electronics*. IEEE, pp.2029 - 2036.