iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45778

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Deep Convolutional Neural Networks for Environmental Sound Classification

Aakash[1], Abhishek S Nigudgi[2], Ankush M Awanty[3], Dr. Suvarna Nandya[4]
[1, 2, 3, 4] *Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering,*

Abstract: ***We propose a model to classify environmental sounds such as People Sounds, Vehicles Sounds, Siren Sounds, Horn, Engine Sounds. We perform Data Augmentation techniques to extract best features from the given audio to classify which class of sound. Our deep convolutional neural network architecture uses stacked convolutional and pooling layers to extract high-level feature representations from spectrogram-like features from the given input.***
Keywords: Data Augmentation, Deep Convolutional Neural Networks, Spectrogram

I. INTRODUCTION

In recent years, the research on environmental sound classification, which is dedicated mainly to identify specific sound events, such as identifying People, Vehicles, Sirens, Horn, Engine sounds has received increasing attention. The Environmental sounds contains so much noise and many sounds which are nowhere related to environment in short it has a lot of disturbance, To deal and classify such sounds we used Deep Convolutional Neural Networks which is one of a machine learning technique.
Classifying environmental sounds that is audio surveillance by the means of CCTV can help monitor better compared to the traditional method in which only video is classified.
Hence summarizing, our project aims to monitor audio in CCTV cameras which helps in improving the surveillance mode.

II. LITERATURE SURVEY

Summary of research papers survey are listed down, the observations made in each paper are summarized below the respective paper and these observations are used to improve the overall system.

1) *Raluca Mus˘aloiu-E S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," IEEE Trans. on Audio, Speech, and Language Processing, Aug 2009*

A variety of features have been proposed for audio recognition, including the popular Mel-frequency cepstral coefficients (MFCCs) which describe the audio spectral shape, Environmental sounds, such as chirpings of insects and sounds of rain which are typically noise-like with a broad flat spectrum, may include strong temporal domain signatures. There are only few temporal-domain features have been developed to characterize such diverse audio signals previously. Here, they perform an empirical feature analysis for audio environment characterization and propose to use the matching pursuit (MP) algorithm to obtain effective time–frequency features. The MP-based method utilizes a dictionary of atoms for feature selection, resulting in a flexible, intuitive and physically interpretable set of features. The MP-based feature is adopted to supplement the MFCC features to yield higher recognition accuracy for environmental sounds. Extensive experiments are conducted to demonstrate the effectiveness of these joint features for unstructured environmental sound classification, including listening tests to study human recognition capabilities. The recognition system has shown to produce comparable performance as human listeners.

2) *J. Salamon, C. Jacoby, and J.P.Bello, "A Dataset and Taxonomy for Urban Sound Research," in 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, USA, Nov. 2014, pp. 1041–1044.*

Automatic urban sound classification can benefit a variety of multimedia applications. In this paper we identified two main barriers to research in this area – the lack of a common taxonomy and the scarceness of large, real-world, annotated data. To address the first issue we presented the Urban Sound Taxonomy, based on previous soundscape research with a focus on sound classes from real noise-complaint data. To address the second issue we presented UrbanSound, a dataset containing 27 hours of audio with 18.5 hours of manually labelled sound occurrences. We also presented UrbanSound8K, a subset of the dataset designed for training sound classification algorithms.

3) *Zohaib Mushtaq, Shun-Feng Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation", Applied Acoustics, Volume 167, 2020, 107389, ISSN 0003-682X*

The adoption of the environmental sound classification (ESC) tasks increases very rapidly over recent years due to its broad range of applications in our daily routine life. ESC is also known as Sound Event Recognition (SER) which involves the context of recognizing the audio stream, related to various environmental sounds. Some frequent and common aspects like non-uniform distance between acoustic source and microphone, the difference in the framework, presence of numerous sounds sources in audio recordings and overlapping various sound events make this ESC problem much complex and complicated. This study is to employ deep convolutional neural networks (CNN) with regularization and data enhancement with basic audio features that have verified to be efficient on ESC tasks. In this study, the performance of DCNN with max-pooling (Model-1) and without max-pooling (Model-2) function are examined. Three audio attribute extraction techniques, Mel spectrogram (Mel), Mel Frequency Cepstral Coefficient (MFCC) and Log-Mel, are considered for the ESC-10, ESC-50, and Urban sound (US8K)datasets. Furthermore, to avoid the risk of overfitting due to limited numbers of data, this study also introduces offline data augmentation techniques to enhance the used datasets with a combination of L2 regularization. The performance evaluation illustrates that the best accuracy attained by the proposed DCNN without max-pooling function (Model-2) and using Log-Mel audio feature extraction on those augmented datasets. For ESC-10, ESC-50 and US8K, the highest achieved accuracies are 94.94%, 89.28%, and 95.37% respectively. The experimental results show that the proposed approach can accomplish the best performance on environment sound classification problems.

4) *R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'05), New Paltz, NY, USA, Oct. 2005*

The paper proposed a hybrid audio analysis framework for surveillance that consists two parts; one that performs unsupervised audio analysis and another that performs analysis using an audio classification framework. The audio classes for the classification framework are obtained from off-line time series analysis of cepstral features and training. It also adaptively learns a Gaussian Mixture Model(GMM) to model the background sounds and updates the model incrementally as new audio data arrives and has been shown to detect suspicious events effectively. The adaptive background modelling algorithm used in the proposed framework first estimates a GMM from WS observations and then updates the parameters of statistically equivalent components in the background GMM. An alternative approach proposed which is analogous to background modelling in computer vision, updates the background model for every new incoming data vector.

5) *C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of lowcost urban acoustic monitoring devices," Applied Acoustics, vol. In Press, 2016.*

The urban sound environment of New York City (NYC) can be, amongst other things: loud, intrusive, exciting and dynamic. As indicated by the large majority of noise complaints registered with the NYC 311 information/complaints line, the urban sound environment has a profound effect on the quality of life of the city's inhabitants. To monitor and ultimately understand these sonic environments, a process of long-term acoustic measurement and analysis is required. The traditional method of environmental acoustic monitoring utilizes short term measurement periods using expensive equipment, setup and operated by experienced and costly personnel. In this paper a different approach is proposed to this application which implements a smart, low-cost, static, acoustic sensing device based around consumer hardware. These devices can be deployed in numerous and varied urban locations for long periods of time, allowing for the collection of longitudinal urban acoustic data. The varied environmental conditions of urban settings make for a challenge in gathering calibrated sound pressure level data for prospective stakeholders. This paper details the sensors design, development and potential future applications, with a focus on the calibration of the devices Microelectromechanical systems (MEMS) microphone in order to generate reliable decibel levels at the type/class 2 level.

6) *R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'05), New Paltz, NY, USA, Oct. 2005, pp. 158–161*

We proposed a time series analysis based approach for systematic choice of audio classes for detection of crimes in elevators, Since all the different sounds in a surveillance environment cannot be anticipated, a surveillance system for event detection cannot completely rely on a supervised audio classification framework. In this paper, we propose a hybrid solution that consists two parts; one that performs unsupervised audio analysis and another that performs analysis using an audio classification framework obtained from off-line analysis and training.

The proposed system is capable of detecting new kinds of suspicious audio events that occur as outliers against a background of usual activity. It adaptively learns a Gaussian Mixture Model(GMM) to model the background sounds and updates the model incrementally as new audio data arrives. New types of suspicious events can be detected as deviants from this usual background model. The results on elevator audio data are promising.

7) *J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, Aug. 2015, pp. 714– 718.*

Hearing aids are increasingly essential for people with hearing loss. For this purpose, environmental noise estimation and classification are some of the required technologies. However, some noise classifiers utilize multiple audio features, which cause intense computation. In addition, such noise classifiers employ inputs of different time lengths, which may affect classification performance. Thus, this paper proposes a model architecture for noise classification, and performs experiments with three different audio segment time lengths. The proposed model attains fewer floating-point operations and parameters by utilizing the log-scaled mel-spectrogram as an input feature. The proposed models are evaluated with classification accuracy, computational complexity, trainable parameters, and inference time on the UrbanSound8k dataset and HANS dataset. The experimental results showed that the proposed model outperforms other models on two datasets. Furthermore, compared with other models, the proposed model reduces model complexity and inference time while maintaining classification accuracy. As a result, the proposed noise classification for hearing aids offers less computational complexity without compromising performance.

8) *"Feature learning with deep scattering for urban sound analysis," in 2015 European Signal Processing Conference, Nice, France, Aug. 2015.*

In this paper we evaluate the scattering transform as an alternative signal representation to the mel-spectrogram in the context of unsupervised feature learning for urban sound classification. We show that we can obtain comparable (or better) performance using the scattering transform whilst reducing both the amount of training data required for feature learning and the size of the learned codebook by an order of magnitude. In both cases the improvement is attributed to the local phase invariance of the representation. We also observe improved classification of sources in the background of the auditory scene, a result that provides further support for the importance of temporal modulation in sound segregation.

9) *C. Bauge, M. Lagrange, J. And N, and S. Mallat,"Representing environmental sounds using the separa-ble scattering transform," in IEEE ICASSP, Vancouver,Canada, May 2013, pp. 8667–8671.*

In this paper we propose a novel representation of such sounds based on the scattering transform which has the property of stability to time-warping deformations and invariance to time-shift useful for classifications tasks. This representation is compared to several state-of-the-art approaches for the task of quantifying similarity between environmental sounds. what we hear. For environmental sounds, everyday listening seems the most appropriate type of listening. How can we design a representation that would be useful for implementing this mode of listening? What kind of acoustic features should be kept in the representation and which should be discarded? To provide some answers to those questions, we propose a new computational approach based on the scattering transform that has the property of time-shift invariance. We review the state-of-the-art in terms of sound representations and similarity computation. We also present the scattering transform and introduces a variant that adds frequency transposition property. We also describe the experimental protocol used to compare the representations.

10) *E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in 2015 International Joint Conference on Neural Networks (IJCNN), July 2015, pp. 1–7.*

In this paper, the use of multi label neural networks are proposed for detection of temporally overlapping sound events in realistic environments. Real-life sound recordings typically have many overlapping sound events, making it hard to recognize each event with the standard sound event detection methods. Frame-wise spectral-domain features are used as inputs to train a deep neural network for multi label classification in this work. The model is evaluated with recordings from realistic everyday environments and the obtained overall accuracy is 63.8%. The method is compared against a state-of-the-art method using non-negative matrix factorization as a pre-processing stage and hidden Markov models as a classifier. The proposed method improves the accuracy by 19% percentage points overall.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

III. METHODOLOGY

In this methodology, we first take original audio data as input and process it for augmentation of the audio data to improve the stability and accuracy of our prediction. In audio augmentation process we alter the audio data by shifting pitches, trimming silence, stretching time quickly/slowly , adding white noises etc. After data augmentation is processed it is now sent to extract features from audio. The process of extracting features will be done with the help of Histograms and spectrograms where we collect data points and plot the respective graphs then we send the produced graphs to CNN networks to train our model, and at the end after the model is trained, the CNN extracts the features and produces the output which specifies the classification of the audio provided belongs to which class. The input we will provide will be the original audio data from real world and not from the already existing data. The architecture design of the model is as shown in the below figure.

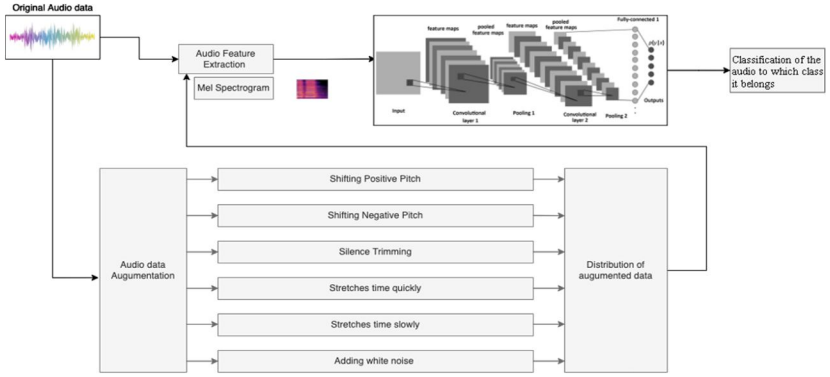


Fig Methodology

IV. RESULT AND DISCUSSION

In this section, the performance of proposed deep convolutional neural networks based classification of environmental sounds system is compared with effectiveness of proposed approach is observed which further evaluated with reference to different parameters. Deep-CNN trained on the recorded dataset. Table1 shows the results comparison of some previous studies conducted for sound classification using spectrogram features. It is seen from table1 that CNN provides better performance compared to other methods.

Spectrogram driven sound	Techniques/ Architectures used	Performance (Accuracy %)
MFCC-SVM	Support Vector Machine (SVM)	34.1%
MPEG-7	Decision Tree	33.6%
Gabor	Random Forest	39.0%
GTCC	K-Nearest Neighbor	40.8%
MFCC-MP	Multilayer Perceptron	43.24%
CNN	Convolutional Neural Networks	73%
TDSN	Tensor Deep Stack Network	56%

Table1: Performance Comparison of different approaches

With more samples of data we train our model, the more the model will provide accuracy. This is because of the ability of DCNN architecture to learn more features from the large datasets. Figure1 shows the accuracy of per class after training the model. As shown People sounds has more accuracy compared to all the other classes as we have trained more number of samples in that particular class, therefore achieving a higher accuracy. Similarly horn class was trained with the less number of samples thus having lower accuracy.

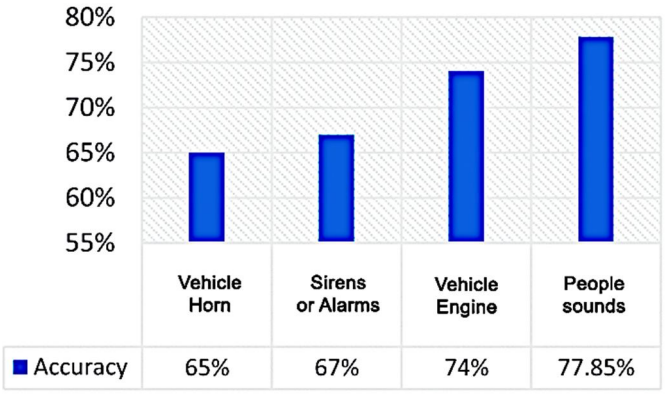


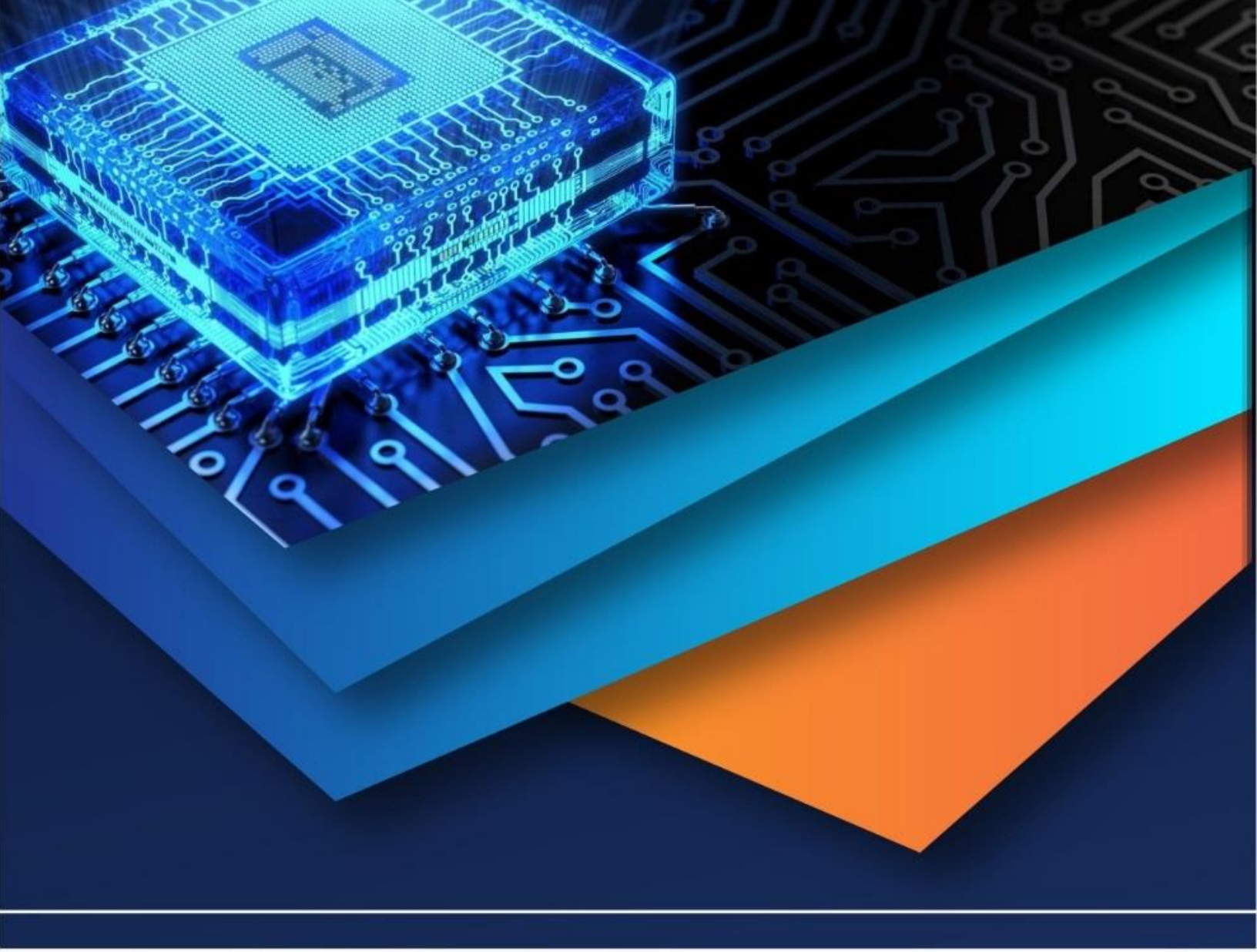
Figure1: Accuracy of classes

V. CONCLUSION

In this we proposed a deep convolutional neural network architecture which, in combination with a set of audio data augmentations, produces state-of-the-art results for environmental sound classification specifically people sounds, vehicles sounds, train sounds, horn, engine sounds. We will show that the improved performance stems from the combination of a deep, high-capacity model and an augmented training set: this combination outperforms both the proposed CNN without augmentation and a shallow dictionary learning model with augmentation.

REFERENCES

[1] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142–1158, Aug. 2009.
[2] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in IEEE Worksh. on Apps. of Signal Processing to Audio and Acoustics (WASPAA'05), New Paltz, NY, USA, Oct. 2005, pp. 158–161.
[3] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of lowcost urban acoustic monitoring devices," Applied Acoustics, vol. In Press, 2016.
[4] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 151–155.
[5] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 6450– 6454.
[6] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 6445–6449.
[7] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 171–175.
[8] "Feature learning with deep scattering for urban sound analysis," in 2015 European Signal Processing Conference, Nice, France, Aug. 2015.
[9] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, Aug. 2015, pp. 714– 718.
[10] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in 2015 International Joint Conference on Neural Networks (IJCNN), July 2015, pp. 1–7.
[11] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, Sep. 2015, pp. 1–6.
[12] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2013, pp. 1–4.
[13] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.
[14] S. Sigtia, A. Stark, S. Krstulovic, and M. Plumbley, "Automatic environmental sound recognition: Performance versus computational cost," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, no. 99, pp. 1–1, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY