FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 20, Nº 3, 2021, pp. 155 - 165
https://doi.org/10.22190/FUACR211015012D

Regular Paper

URBAN SOUND RECOGNITION USING DIFFERENT FEATURE EXTRACTION TECHNIQUES

UDC ((613.644+613.164):004.85:004.032.26)

Simona Domazetovska, Viktor Gavriloski, Maja Anachkova, Zlatko Petreski

University Ss Cyril and Methodius in Skopje, Faculty of Mechanical Engineering in Skopje, Republic of North Macedonia

Abstract. *The application of the advanced methods for noise analysis in the urban areas through the development of systems for classification of sound events significantly improves and simplifies the process of noise assessment. The main purpose of sound recognition and classification systems is to develop algorithms that can detect and classify sound events that occur in the chosen environment, giving an appropriate response to their users. In this research, a supervised system for recognition and classification of sound events has been established through the development of feature extraction techniques based on digital signal processing of the audio signals that are further used as an input parameter in the machine learning algorithms for classification of the sound events. Various audio parameters were extracted and processed in order to choose the best set of parameters that result in better recognition of the class to which the sounds belong. The created acoustic event detection and classification (AED/C) system could be further implemented in sound sensors for automatic control of environmental noise using the source classification that leads to reduced amount of required human validation of the sound level measurements since the target noise source is evidently defined.*

Key words: *Sound recognition, audio parametrization, machine learning, urban noise*

1. INTRODUCTION

Based on the advanced technology of artificial intelligence (AI), acoustic event detection and classification systems are developed for simplifying the classic traditional methods for estimation of the noise pollution [1]. The main objective of the acoustic AI-

Received October 15, 2021 / Accepted November 17, 2021
Corresponding author: Simona Domazetovska
University Ss Cyril and Methodius in Skopje, Faculty of Mechanical Engineering in Skopje, Karpoš II bb, 1000 Skopje, Republic of North Macedonia
E-mail: simona.domazetovska@mf.edu.mk

based systems is to develop algorithms able to recognize and classify the sound events in classes that are appropriate for the chosen acoustic environment.

A key requirement for audio signal classification is the extraction of appropriate acoustic parameters that represent the important audio features of the signal. The audio feature extraction is used for extracting and understanding meaningful information from audio signals in order to obtain more expressive and compact visualization of the signal properties. Analyzing the different audio parameters and comparing their contribution in the classification accuracy could result in choosing the right set of parameters that help in establishing high recognition of the classes of sound events.

The AED/C systems can find their use in many IoT-based applications for development of smart cities for urban noise classification. According to the project in [2], the researchers have established dynamic noise maps based on an anomalous noise events detector that was initially designed and trained using data from a real-life recording campaign, where several audio parameters were chosen to improve the system accuracy. Furthermore, the SONYC project has deployed 56 low-cost acoustic sensors across New York City to monitor the urban noise and perform a multi-label classification of urban sound sources in real time [3]. Another application-oriented example on AED/C is the development of an embedded device where the authors present an acoustic approach to emergency vehicle detection (e.g. ambulances, police cars) through the detection of the sound generated by their sirens [4].

Even though the content of environmental sounds is more diverse than speech and music signals, the features established for acoustic sound recognition and music instrument retrieval are widely used in environmental sound classification due to their significant performance. The researchers in the study [5] analyze the performance by using aggregated acoustic features for environmental sound recognition by using deep learning classifier. By using several feature extraction techniques as an input to machine learning algorithm, the researchers in [6] have studied which set of features will give the highest efficiency of the system when classifying urban noise. The approach in this study uses MFCC for audio feature extraction and supervised classification algorithms (SVM, KNN, Bagging, and Random Forest) for noise identification. By using several datasets of environmental sound classes, the researchers in [7,8] have examined and visualized the MEL spectrogram feature as an input in the CNN network. As it could be noticed, most of the prominent and traditional audio features extraction techniques used for environmental classification are MFCC, Mel Spectrogram and Wavelet features. For urban sound classification systems, the acoustic features are mostly modeled by using the Gaussian Mixture Models (GMMs) [9], Support Vector Machines (SVMs) [10], Hidden Markov Models (HMMs) [11] and the neural network-based approaches, such as DNN [12] and CNN [13].

Inspired by the previous studies, the motivation in this paper is to examine several feature extraction techniques that are widely used and combine them to analyze their efficiency when using traditional ML algorithms. By investigating the performance of the features with three different machine learning classifiers (Random Forest, Support Vector Machines and Naïve Bayes classifier) that are less used for the purpose of urban noise classification but known to show good classification results in other audio classification systems, novel classification system will be proposed and tested.

In this paper, a supervised system for acoustic event detection and classification based on the AI techniques will be analyzed, describing the system design, and applying various audio parameters and ML algorithms in order to form a system that will predict sound events to the class where they belong with high accuracy. The designed system is

used for training and testing labeled data which includes 10 classes of disturbing urban sound events. The focus of this paper is on the feature extraction processes, analyzing the accuracy that can be established using three audio parameters that are widely used in the field of sound recognition: MEL Frequency Cepstral Coefficients (MFCC), MEL Spectrogram and Chromagram. The extracted features from the used audio parameters are merged into a single feature vector that represents the sound signal, which is subject to recognition, and is further used as an input in the chosen ML algorithms.

The organization of this paper is as follows: Section 1 shows and describes the architecture of the supervised AED/C system that will be used for the purpose of this research. In addition, the dataset of the urban sounds used for training and testing the system will be analyzed. Section 2 defines the audio parameters used for extracting the features, focusing on the used digital signal processing steps, while section 3 provides the accuracy results of the implemented AED/C system. At the end, in section 4, the conclusions and the proposed future work are discussed.

2. DESIGN OF AED/C SYSTEM

The used architecture for designing the supervised system for detection and classification of sound events is shown on Figure 1. To design the AED/C system, three main processes including detection, feature extraction and classification must be applied. The system must first go through a training process using a database with known sound events to create the acoustic model, and afterwards the testing process is applied using unknown sound events, so that the accuracy of the system could be validated.

In the detection part, as the first and very important step for high accuracy establishment, it is essential for each sound event to be systematized according to the proposed sound taxonomy. Based on the proposed taxonomy, a database of selected sound events in the chosen environment is created. Each sound event that takes part of the database needs to be pre-processed in order to create labeled data that can be used for supervised learning methods. The processed audio signals form a database that can be used as input in the AED/C systems.

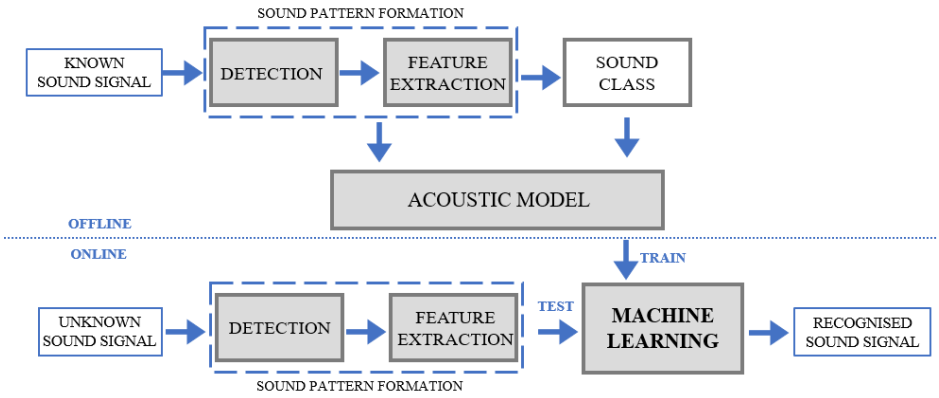

Fig. 1 Architecture of system for detection and classification of sound events

In the first phase, the sound events are detected from a continuous sound that is recorded with a microphone. Sound event detection enables segmentation and separation of the sound events of interest and can be done using two approaches: detection and classification (using event detection system defined as any rapid change in long-term background noise) and detection by classification (classification of each segment in a fixed length of time). The purpose of the sound detection process is to detect the sound event by finding its starting and ending points.

The next step is the feature extraction process, where the features of the sound segment are extracted and compactly represented using the audio parameterization process. The sound signal is divided into short time frames, usually between 10-50 milliseconds. This process has a dual application, on the one hand, the non-stationary audio signal is considered to be stationary for each defined short time frame, enabling the spectrum-time analysis, and on the other hand increases the efficiency of the process itself. For the purpose of this research, the signal was framed into 20ms windows because of its sufficiency to provide a good spectral resolution of the sounds and at the same time is short enough to resolve a significant temporal characteristic. Afterwards, the digital signal processing techniques are applied to extract audio features from the audio signal. The extracted features form a feature vector, which is a single vector that represents the sound signal as an object for recognizing certain features. The process of connecting the features in a single vector can cause a problem while processing the data in the machine learning algorithm through the appearance of high dimensionality of the vector itself. The feature extraction is followed by a process of reducing the dimensionality of the data so compact feature vectors could be obtained. The features determine which audio properties are available and will be processed, while the information that is not extracted by the features is unavailable to the system. The selection of appropriate signal features is key issue for successful environmental sound recognition. For this research, three audio parameters were chosen while building the AED/C system.

Finally, a supervised machine learning algorithm is applied for training and testing the system. The system must first go through a training process known as the offline process which creates an acoustic model based on labeled dataset for training the machine learning algorithm. The next step is testing, i.e. recognizing unknown sound signals and classifying them into an appropriate class based on the information gained during training (known as the online process). The used audio parameters were tested using three machine learning algorithms. The applied ML algorithms were used as classifiers for the supervised learning method. Random Forest is a supervised learning algorithm where the built forest is an ensemble of Decision Trees, and by building multiple decision trees and merging them together, a more accurate and stable prediction is made. The Support Vector Machines supervised learning methods analyze data and recognize patterns in multi-class classification by applying different kernel function. The Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independent assumptions between the features. Naive Bayes model is easy to build and particularly useful for very large data sets.

The final goal is to have high classification accuracy of the AED/C systems with sound events that are appropriately classified in the class where they belong.

2.1. Urban sound dataset

The ideal AED/C system based on the machine listening technique should successfully identify certain sound events from a wide variety of sounds. Accordingly, it is necessary to limit the number of sounds for analysis depending on the type of sound and the need for classification. As the sound systematization is the first step when designing the system, an audio taxonomy needs to be developed to systematize sound that helps in better understanding the domain of the data being processed. By dividing the sound events into appropriate classes, the accuracy of the classification could be improved. The systematization of a sound event regardless of the nature of the sound is a complex problem, so the purpose is to divide the sound into simpler and smaller classes of events for easier recognition when using the machine learning algorithms.

The sounds of interest in this research are the sounds causing the urban noise pollution, as they have a negative influence on the citizens health. Based on the numerous studies, the researchers in [14] have formed a dataset based on a defined taxonomy for the urban sound research. Due to the proposed requirements, the taxonomy for the urban sound annoyance is divided in 10 classes of sound events: engine idling, jackhammer, car horn, gun shot, children playing, drilling, street music, dog bark, air conditioner and siren. The dataset UrbanSound8K that will be used in this research was created based on the previously formed taxonomy, and it contains 8732 audio signals with the total length of around 8 hours, designed for training and testing machine learning algorithms. Each sound event has duration of maximum 4 seconds that is enough time to identify the sound event class.

3. FEATURE EXTRACTION

The development of systems for acoustic event detection and classification is a method comprised of processing acoustic signals and converting them into symbolic descriptions that correspond to the various sound events which describe each of the analyzed audio signals [15]. The amount of raw data in the signal is too big for direct processing, that is why feature extraction aims at reducing the amount of data and extracting meaningful information from the signal. The result of the feature extraction process are parametric numerical features that characterize meaningful information of the input signals. The extracted features are stored in the feature database and are used for feeding the machine learning algorithm.

The proper selection of audio features for the sound events from the used dataset has a key role in achieving a successful AED/C system. Based on the previous studies, three audio parameters known to be effective when classifying urban noise were chosen: MEL Frequency Cepstral Coefficients (MFCC), MEL Spectrogram and Chromagram.

The MFCC parameter will be used as base audio parameter, as it has been largely employed in the field of environmental sound classification, especially in the field of music and environmental sound classification. The steps for extracting MFCC are shown in Figure 2. Firstly, a Discrete Fourier transform of the windowed input signal is computed. Then a Mel-filter bank based on a perceptual-based frequency scale (human auditory model in which the MEL-scale frequency is inspired), consisting of logarithmically positioned triangular band-pass filters is applied. After taking the logarithm of the magnitude of the band-pass filtered amplitudes, the Cosine transform results in obtaining MFCCs.

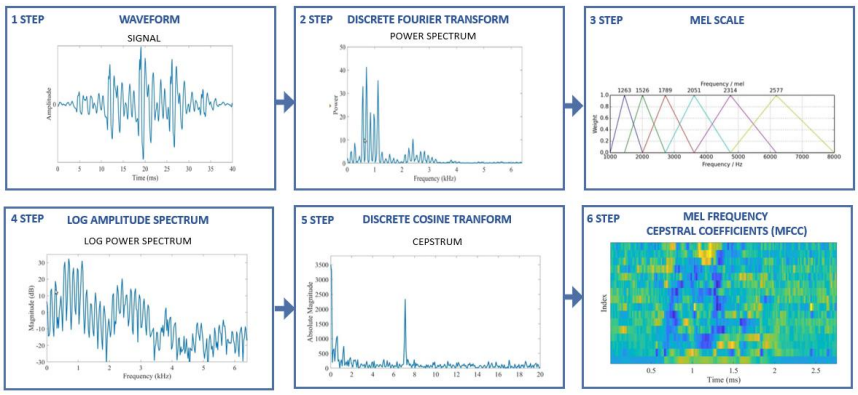

Fig. 2 MFCC feature extraction steps

The steps can be explained by a signature for MFCCs by selecting the necessary building blocks from transformations, filters, aggregations and detectors proposed in [12]. First, a single frame ("f") of the input signal is extracted and a Discrete Fourier transform ("F") is performed. Then spectral binning of the Fourier coefficients is performed to obtain the responses of the Mel-filters ("β"). Taking the logarithm corresponds to "l" and the completing Cosine transform matches "C". The resulting sequence for the MFCC feature is "f F β l C".

The MEL spectrogram is the visual representation of the spectrum of frequencies of sound signal as they vary with time based on the MEL-scale. This audio parameter represents the signal in time and frequency domain, obtaining the optimal balance which allows accurate representative signals. The steps for extracting the MEL spectrogram feature are shown in Figure 3. First, the audio signal is converted from time to frequency domain using the fast Fourier transform. By converting the frequency y-axis to a log scale and the amplitude into color dimension, a spectrogram is formed. Applying the MEL scale on the frequency axis results into forming the MEL spectrogram which allows better understanding of the processed image. The MEL-spectrogram shows high accuracy in the ML systems when classifying the sound events, especially when using algorithms based on deep learning.

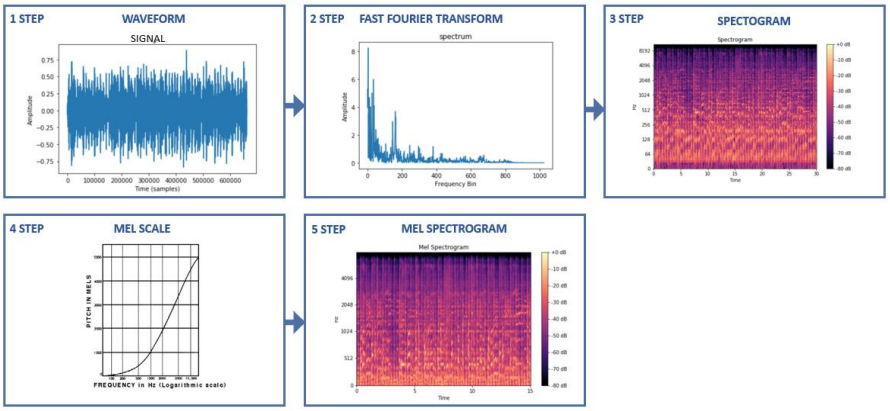

Fig. 3 MEL Spectrogram feature extraction steps

The Chromagram feature is related to the perception of pitch in a sense that it is a complement of the tone height. Chromagram or chroma-based feature is a spectrum-based energy representation that considers the 12 pitch classes within an octave (corresponding to pitch classes in musical theory), and it can be computed from a logarithmic short-time Fourier transform [16]. Figure 4 shows the needed digital signal processing steps for visualization and extracting the Chromagram feature, so the time-frequency properties transform into a temporally varying precursor of pitch. The transformation is based on perceptual observation concentrating on the humans' auditory system.

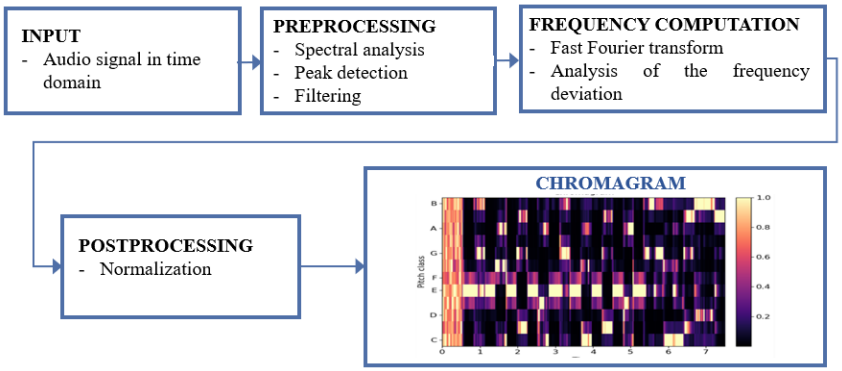

Fig. 4 Chromagram feature extraction steps

The Chromagram feature seems to provide a more direct measure of the variations related to pitch and give higher accuracy in the prediction than the MEL spectrogram feature.

4. RESULTS

A system for acoustic event detection and classification was built and tested using the above-mentioned features and ML algorithms. While testing and training the system, nine folders of the database were used for training, while one folder was used for testing the model. In the first phase, the MFCC audio parameter was used as the base parameter as it gives high accuracy of the system and have been largely employed in the field of environmental sound classification. The first few coefficients of MFCC describe the rough spectral shape; the first coefficient represents the average power in the spectrum, the second one the spectral centroid etc. Normally, the first 20 coefficients are used to represent the shape of the spectrum, but some applications need to use higher-order coefficients to extract as much information as possible [17]. Because this research deals with various sound events from different domain that are generated in the urban environment, their variation makes the recognition more complex, requiring higher number of coefficients. That is why the MFCC parameter was tested with different number of coefficients, varying from 10 to 50. Table 1 shows the results from the established accuracy when testing the AED/C system with all the three ML algorithms. As shown on the table, the highest accuracy happens when using 40 coefficients of the MFCC parameter. The highest accuracy of 55.07% was

established for the RF classifier, afterwards the SVM with 51.05% accuracy and in the end 47.19% for the NB classifier. Because the highest achieved accuracy for all the three machine learning classifiers happened when using 40 MFCC coefficients, it was decided to use this number of coefficients when extracting the MFCC parameter and combining it with the other two audio features.

Table 1 Established accuracy from the three ML algorithms for different number of MFCC coefficients

	Number of MFCC coefficients				
	10	20	30	40	50
RF	30.12 %	48.02 %	50.13 %	55.07 %	55.02 %
SVM	27.72 %	47.12 %	49.02 %	51.05 %	50.85 %
NB	25.61 %	44.52 %	46.3 %	47.19 %	47.19 %

Next, the extracted MFCC parameters with 40 coefficients will be combined using the MEL Spectrogram and Chromagram audio features. The feature vector for each audio file will be formed using different combinations of the chosen audio parameters in order to create a system with the feature vector that gives the highest accuracy when distinguishing the classes of the sound events. Figure 5 shows the achieved accuracies, while Table 2 shows the total number of extracted coefficients for each audio parameter.

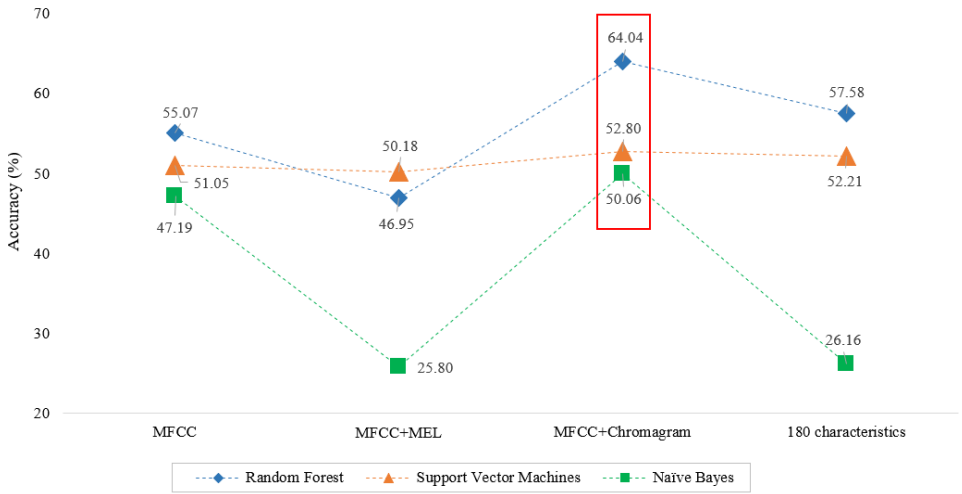

Fig. 5 Accuracy results using different audio parameters and ML algorithms

Table 2 Number of coefficients for each of the audio parameter

Audio parameters	Number of coefficients
MEL Frequency Cepstral Coefficients (MFCC)	40
MEL Spectrogram	128
Chromagram	12

As it can be noticed from the accuracy results shown on figure 5, the MEL spectrogram feature significantly decreases the accuracy of the system. Namely, when using the Naïve Bayes classifier, the system reaches only half of the achieved accuracy compared to the case when only MFCC parameter is applied. It could be noted that the MEL spectrogram decreases the accuracy when using the traditional ML algorithms, although according to the other studies this parameter gives high accuracy results when using the algorithms based on deep learning.

When using the MFCC and the chromagram feature, the accuracy increases for the three ML algorithms, where the highest accuracy of 64.04% is established when using the Random Forest classifier, while for the other two classifiers the accuracy increases up to 3%.

Combining the three audio parameters which form the feature vector with 180 coefficients (40 MFCC, 128 MEL Spectrogram and 12 Chromagram) results in less accurate results than when only the MFCC and Chromagram audio parameters are used.

The MEL spectrogram as a visual tool can be faulted because the pitch and the resonance in the vocal production are not readily separable in the visual representation of the signal like the chromagram feature is. That's why, the chroma feature seems to provide a more direct measure of the variations related to pitch and give higher accuracy in the prediction when classifying the urban noise.

According to the research in [18], the accuracy of designed AED/C system for the UrbanSound8K dataset was investigated using several audio parameters and machine learning algorithms. In comparison to their study, the achieved accuracy in this paper is greater for the Random Forest classifier with 3% higher results, and almost twice higher accuracy for the Naïve Bayes classifier, while for SVM the performance is slightly lower. The contribution in this research with respect with the previous studies shows different AED/C architecture by using different feature extraction techniques as input in traditional ML algorithms that result in higher accuracy results than previous studies. It can be concluded that the combination of the MFCC and Chromagram gives better prediction results that when combining the MEL spectrogram. The overall accuracy of the tested results could be improved by applying the hyperparameter optimization on the ML algorithms.

When analyzing the results, it can be noticed that the sound event representing the 'engine idling' has the highest classifying accuracy, while for the 'street music' the highest errors could be noticed. Analyzing the visual representations from the audio parameters, it could be noticed that the 'engine idling' sound event has similar images than the 'street music'. This could be due to the sound event because the street music has many elements and types of the sound event, while for the engine idling it could be stated out that this sound event has similar visual representation for each of the audio file that represents this event. Figure 6 shows visual representation of the extracted audio parameters for the sound class of 'engine idling' and 'street music'.

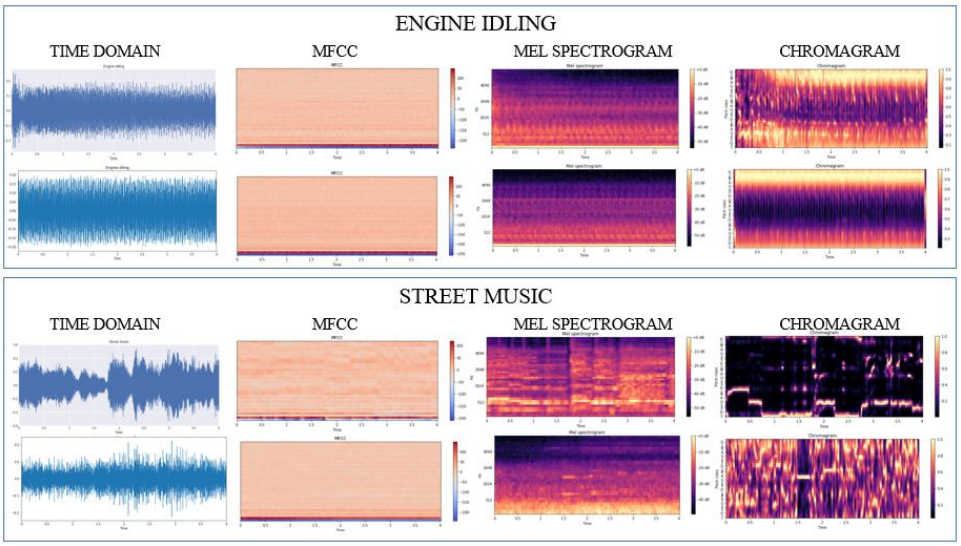

Fig. 6 Visual representation of the used audio parameters of the sound events 'engine idling' and 'street music'

5. CONCLUSIONS

The focus in this paper was on the feature extraction techniques, where three audio parameters were used and combined in order to choose the best set of parameters that achieves the highest accuracy when testing the AED/C system. The best accuracy result was established by combining the MFCC and Chromagram audio parameters forming a feature vector consisting of 52 coefficients (40 of MFCC and 12 of chromagram). Using the Random Forest classifier resulted in achieving the best results with 55.07% accuracy. The MEL spectrogram feature confuses the prediction by showing low accuracy results. From here, it could be stated that this parameter is not suitable for urban noise recognition when using the traditional ML algorithms.

From this research, we can conclude that the use of the two audio parameters: MFCC and Chromagram have shown the best results when designing AED/C system for the chosen urban sound dataset, but still, improvement is needed. The achieved accuracy is not satisfactory to enable practical application of the system. The future work of this research proposes focusing on the ML algorithms while using the MFCC and the chromagram feature. By applying hyperparameter optimization on the machine learning algorithms, the designed AED/C system could result with high predictive accuracy results so it would be suitable for practical application.

REFERENCES

[1] Socoró, J. C., Sevillano, X., & Alías, F. *Analysis and automatic detection of anomalous noise events in real recordings of road traffic noise for the LIFE DYNAMAP project*. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings (Vol. 253, No. 6, pp. 1943-1952),. Institute of Noise Control Engineering. (August, 2016). Available: https://doi.org/10.1515/noise-2018-0006

[2] Socoró, J. C., Alías, F., Alsina R. M, Sevillano X., Q. B3-*Report describing the ANED algorithms for low and high computation capacity sensors.* 2016

[3] Bello, J. P., Silva, C., Nov, O., Dubois, R. L., Arora, A., Salamon, J. Sonyc*: A system for monitoring, analyzing, and mitigating urban noise pollution.* Communications of the ACM, 62(2), 68-77. 2019. Available: https://doi.org/10.1145/3224204

[4] Hollosi, D., Nagy, G., Rodigast, R., Goetze, S., & Cousin, P.. *Enhancing wireless sensor networks with acoustic sensing technology: use cases, applications & experiments*. In 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE CPS Computing (pp. 335-342). IEEE. (2013, August*)* Available: 10.1109/GreenCom-iThings-CPSCom.2013.75

[5] Su, Y., Zhang, K., Wang, J., Zhou, D., & Madani, K. Performance analysis of multiple aggregated acoustic features for environment sound classification. *Applied Acoustics*, *158*, 107050. (2020) Available: https://doi.org/10.1016/j.apacoust.2019.107050

[6] Alsouda,Y., Pllana,S., Kurti, A.. *Iot-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest.* In Proceedings of the international conference on omni-layer intelligent systems (pp. 62-67). (2019, May*)* Available: https://doi.org/10.1145/3312614.3312631

[7] Zhang, Zhichao, et al. *"Learning attentive representations for environmental sound classification." IEEE Access* 7 (2019): 130327-130339.Fu Z., Lu, G., Ming Ting, K., Zhang, D. *A survey of audio-based music classification and annotation [Journal] // IEEE Transactions on Multimedia*. 2:Vol.13. - pp. 303-319. (April 2011) DOI: 10.1109/TMM.2010.2098858

[8] Mushtaq, Z., Su, S. F., & Tran, Q. V. *Spectral images based environmental sound classification using CNN with meaningful data augmentation.* Applied Acoustics, 172, 107581. (2021) Available: https://doi.org/10.1016/j.apacoust.2020.107581

[9] Geiger, J., & Helwani, K. *Improving event detection for audio surveillance using Gabor filterbank features.* In 23rd European Signal Processing Conference (pp. 714–718). (2015) Available: 10.1109/EUSIPCO.2015.7362476

[10] Mulimani, M., & Koolagudi, S. G. *Segmentation and characterization of acoustic event spectrograms using singular value decomposition*. Expert Systems with Applications, 120 413–425. (2019) Available: https://doi.org/10.1016/j.eswa.2018.12.004

[11] J. Schroder, B. Cauchi, M. R. Schadler, N. Moritz, K. Adiloglu, J. Anemuller, S. Doclo, B. Kollmeier, and S. Goetze, *"Acoustic event detection using signal enhancement and spectro-temporal feature extraction,"* in IEEE Workshop on Applicat. Signal Process. Audio Acoust.(WASPAA), pp. 1–3. 2013

[12] Liu, Chengwei, et al. *"Environmental Sound Classification Based on Stacked Concatenated DNN using Aggregated Features."* Journal of Signal Processing Systems 1-13 (2021). Available: https://doi.org/10.1007/s11265-021-01702-x

[13] 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) Available: https://doi.org/10.1016/j.asoc.2019.105912

[14] Salamon, J., Jacoby, C., & Bello, J. P. *A dataset and taxonomy for urban sound research*. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 1041-1044). (November, 2014*)* Available: https://doi.org/10.1145/2647868.2655045

[15] Temko, A., Nadeu, C., Macho, D., Malkin, R., Zieger, C., & Omologo, M.. *Acoustic event detection and classification*. In Computers in the human interaction loop (pp. 61-73). (2009*)*. Springer, London

[16] Mitrović, D., Zeppelzauer, M., & Breiteneder, C. *Features for content-based audio retrieval. In Advances in computers* (Vol. 78, pp. 71-150). Elsevier. (2010) Available: https://doi.org/10.1016/S0065-2458(10)78003-7

[17] K. Wang and C. Xu. *Robust soccer highlight generation with a novel dominant-speech feature extractor*. In Proceedings of the IEEE International Conference on Multimedia and Expo, volume 1, pages 591–594, Taipei, Taiwan, Jun. 2004. IEEE, IEEE. 10.1109/ICME.2004.1394261

[18] Chang, C., & Doran, B. *Urban Sound Classification: With Random Forest SVM DNN RNN and CNN Classifiers*. In CSCI E-81 Machine Learning and Data Mining Final Project Fall 2016. Harvard University Cambridge. 2016