**Title of Research:**

An Extensive Analysis on different features extraction techniques and Machine Learning Models on the Sound Signals for the Automatic Sound Classification Systems (ASCS)

**Objectives and Motivation:**

Sound Signals are one of the unstructured data sources which have not been heavily explored the way the other datasets are explored. The automatic speech recognition devices which are ASR systems, depends heavily on the way the features are extracted. These features extracted forms the basis for the sound signals to be classified for. In this research, the following topics will be analysed in depth:

1. To identify the important features like FFT (Fast Fourier Transform), PSD (Power Spectral Density), AC (Auto Correlation), MFCCs (Mel-Frequency Cepstral Coefficients), Mel Spectrograms, Tonnentz, Contrast and Power Density that will be important for different data sources.
2. Identify the sound, when the human voices are present when other kinds of sounds are present like coughing (CoVID-cases), Ambulance Sounds (Traffic Forecasting for the Traffic) to get the priority of the vehicle movement
3. Analysis of Different Machine Learning Methods like Logistic Regression, Support Vector Machines, Decision Trees, Random Forest and XG-Boost
4. Analyse the Power Density features with the Deep Learning Techniques like Convolutional Neural Networks to find the hidden features that contribute to a particular class

**Introduction:**

Humans classify audio signals all the time without conscious effort. Recognizing a voice on the telephone, telling the difference between a telephone ring and a doorbell ring, these are tasks that we don’t consider very difficult. Problems do arise when the sound is weak or there is noise or itis similar to another sound. For example, I find it difficult to tell which of the doors in the hall outside has just closed. There are three main areas of motivation for ASC research. First, it would be instructive to knowhow it is that humans do what they do. If we knew the general systems that we use to classify audio, we might be able to better diagnose and treat auditory ailments. The research that would answer these questions tends to be more psychological and physiological than computational, but the methods used in computer ASC systems might provide a starting point for human ASC research. Second, it would be nice to have a machine that could do what a human could do with sound. For Example, doctors listen to the way a patient breathes in order to diagnose respiratory ailments, and if a medical expert system could do the same, having been programmed with ASC knowledge, then remote areas could get diagnoses quickly without the expense of consulting a human expert who might be in a different country and need to be transported. In the same way, expert auto mechanics are able to diagnose car engine problems by listening to the sounds that the engine makes as it runs. There are many areas where human experts use their ears in their job. ASC expert systems could provide the opportunity for this work to be done in remote communities or in other situations where an expert would be inaccessible or expensive. Finally, an ASC system has the potential to hear much better than a human. If computers could help us perceive sound in the same way that microscopes, television cameras and instant-replay have helped us to perceive the visual world, we could know much more about the world than we do now. Many tape recorders and voice mail systems provide variable speed playback, allowing the user to fast-forward across easily understood or low content sections, and to slow down and listen carefully to noisy or complicated sections. An ASC system could be constructed to automatically alter the playback speed to keep a constant rate of information. Further potential hearing augmentation applications include noise removal, sound separation and automatic transcription of music, text, Morse code or other sounds used for communication. In daily life, we experience different kind of sounds like birds chirping, music, doorbell, sound of vehicle horns etc. Buses whir and makes sounds as they stop at each section to pick up and drop passengers. Similarly sound generate when the air rumbles while planes fly overhead. Birds chirp, barking dogs, cars honk, and sirens etc. Sound and noise make the daily life of the city quite difficult but we don’t have any escape for it. Large population density creates a lot of human noises and machinery sounds because of day to day tasks, and the services required to keep that population functioning and happy. The analysis and understanding of sound in the urban environment is very important to the growth of cities and populations of the future. Studies have shown a direct impact on health and child development from noise pollution (Stansfeld and Matheson, 2003, Waye et al., 2001, Ising et al., 2004). Here, we have focused on a specific dataset containing sound clips named as URBAN SOUND dataset. Urban sound classification has a lot of scope in the field of nowadays research. Particularly, research in this domain brings new perspective and approaches to support the idea of smart city which will be very helpful for mankind. This can be achieved by improving the quality of existing environments & human activity related sounds and developing better ideas to remove noise generation & filtering out the characteristics of good sonic environments. This project consists of using the sound signals that are one of the major source of unstructured data to extract the information of it and then use various kinds of predictive analysis to see how the sound files falls into the different class. Here we have taken the data from some of the online sources which has different kinds of sounds from 10 different classes. As the sound is a time series data and a kind of longitudinal wave that needs a medium to travel, different sources will have different kinds of frequency and amplitude that will be varying from each other. We have tried to used different kinds of the feature extraction techniques with combination of different machine learning algorithms to find out how the algorithms behave with the features and how the results are.

**Related Work:**

A lot of terminologies are proposed in late 70s. Most of the researchers has their own researched terminologies, but these taxonomies tend to have one basic element on common, which Gustavino estimated with her approach-wise study: the difference between sounds related to the presence or absence of human task. One of the starting terminologies related to environmental sound was given by Schafer in (Schafer et. al, 1977), which distributes sounds at the upper level into six classes: “natural”, “person”, “machinery”, “quiet”, “social”, and “signals”. This terminology forms the base which can be iterated by further terminologies. After Schafer, the Gaver terminology is one of the most preferred terminologies related to environmental sound. However, Gaver’s work elaborates sound events by the things that connects during the task which is the reason for the sound and method for interaction between the things, not in phrases of semantic tasks, as Gustavino clarifies is conceptually relevant. Therefore, the most comprehensive research of “soundscape” and showcase of a related terminology till now is the study of Brown (Brown et al., 2011). Brown’s terminology divides from the root node, “The Sonic Environment”, into “close room” and “outside room”. The terminology is then splitted in a parallel path to Gustavino’s: both divided from “sounds which are created by human tasks” and “sounds not created by human tasks”. Brown’s terminology, further splits sounds in a hierarchical system, where it is divided into “places”, “classes of sound roots”, and “sound roots”. “Sound roots” are the “end points” of the tree, although all of 12 the leaves are not compared to low-level, approach-wise speaking; the end-points “fireworks” or “speech” are very specific than the leaves “rail traffic” or “regeneration”. Therefore, Brown’s terminology is the most significant one we found, and the very much directly related to explaining urban sounds. One last approach to the generation of an environmental sound terminology uses a estimation method (Gygi et al., 2007). This study go through a collection of different type of sounds which are fully dependent on a hierarchical algorithm (ex: clustering) with their sonic signal similarity, and sums up those results with lot of conceptual studies. While this study does not represent a clear terminology, it presents a significant look at both the perceptually related and the sonically related characteristics of different voices or environmental sounds. This section presents the literature related to the feature extraction techniques and machine learning algorithms which are used for sound classification in this study. This research work is an opportunity to excel the state of the art in automatic classification of sonic sounds with developments in feature extraction or new feature generation which have recently been modeled for same kind of audio problems (Coates et al., 2012). Features extracted using fast fourier transform on each signal has the benefit pf reducing the frequency band captured by each signal (Zakaria et al., 2016). Spectral contrast are octave based features which represents spectral properties or relative spectral distribution of an audio signal (Ziang et al., 2002). According to (Muda et al., 2010), MFCCS (Mel Frequency Cepstral Coefiicients) are very significant audio features to perceive human auditory system. (Hibare et al., 2014) stated that feature extraction is a very significant step in signal processing and he presented a study of some feature extraction techniques. He presented Fast fourier transform, MFCCs and Linear predictive coding have great complexity in feature vector dimensions and because of static window size, these techniques are not appropriate for dynamic audio signals like speech or environment sounds etc. He presented some other techniques like wavelet transformation where computational complexity and size of sound feature vector will be reducedwhich will be result in better accuracy. Therefore, wavelet transformation will be good for dynamic signals.

(Jiang et al. 2002) presented that automatic sound classification is very crucial and helpful to manage musical datasets. He presented a study of octave-based features called spectral contrast which was proposed to explain the spectral properties of an audio clip. The experiments in his study showed that octave-based features performed excellently well for audio signal classification. He presented one more comparison that spectral contrast features have better differentiation among different sound signals compare to MFCCs or mel-frequency cepstral coefficients. Coates and Ng had implemented some ML models for feature learning using spherical k-Means, and replicated part of the work done by Dieleman using learned Sphereical k-Means features for audio categorization. In last one decade, machine learning approaches have become very applicable in different fields like medical, sports, finance and it provides very important result (Friedman et al., 2008). He stated that machine learning approaches have the potential to deal with large datasets, reproducing qualitative data regardless of complexity and their capability of predictability and interpretation. Therefore, many studies have used machine learning techniques in audio signal classification to check their explanatory capacity (Martin et al., 2009) and predictive potential (Matias et al., 2008). Naïve Bayes technique makes the learning simple by taking the assumption that all the input features should be independent of each other. Although, independence between input features is not very good assumption, but still Naïve Bayes performs quite well compare to other machine learning techniques in terms of performance and computation time (Rish et al., 2001). Support vector machine, developed by Weston (2014), is a supervised learning based model that is used for regression analysis and classification. This algorithm utilizes the concepts of supervised learning and incremental text mining. This algorithm works based on the structural risk minimization principle from computational theory. Random Forest algorithm, described by Belgiu and Drăguţ, (2016), is capable of performing both jobs: classification of data and regression analysis and it performs better than some other techniques like, k-NN, decision tree algorithms, Naïve Bayes etc. (Joachims et al., 1998) expands the use of Support Vector machine in text classification using some examples for training. Support Vector Machine performs better than other algorithms such as k-NN and Rocchio algorithms. According to (Chen et al., 2016) XGBoost algorithm is represented as an end to end tree boosting model which is very scalable and it results very good performance in many data science use cases. It was developed to push the limits of estimation for boosted trees. It is very much modified and improved version of gradient boosting classifier (GBM). GBM algorithm works sequentially but XGBoost works parallel for trees, which makes it significantly faster. XGBoost is very well scalable and it has memory-bounded settings. This scalability is the result of multiple algorithmic optimizations. It is an approximate algorithm is used to find the best splits when we have to deal with lot of features. Data sorting is performed in parallel among multiple threads of CPU to speed up data-preprocessing part. XGBoost is an algorithm which can deal with sparse data, overfitting (using regularization parameters). Also, it solves cache miss problem and performs out of core calculations. Next chapter will elaborate the dataset used in this study and visualize several feature plots to observe the difference between different sound classes. Further chapters of report are structured as follows: Next section of this chapter presents the objective of our study. Chapter 2 describes the methodology used in current study including the several feature extraction techniques and classification models. Chapter 3 presents the implementation of our study, which explains datasets used for current study: urban sound classification data, performs data visualization to present the difference between different types of sounds using multiple feature extraction techniques and steps used in our study. Chapter 4 puts forward the results and comparative study to find best feature combinations and best classification model among all the features and models used in this study. Chapter 5 presents the summary and findings of present study and scope of future work. References for the project work are presented in the last section.

**Proposed Plan:**

**Dataset Sources:**

**Data Pre -processing:**

**Model Development:**

**Results Evaluation:**

**Please outline the proposed sample group, including any specific criteria:**

As the data used for this project is open source, a series of guidelines and methodologies were followed by the data aggregating organization. While referring to the report, it was mentioned that while doing collection of data variance of data being collected was given highest priority, in which people irrespective of their state, caste, color, race or ethnicity were requested to contribute for a noble project which will benefit the society. It was made sure that we will have equal representation of males and females, as well as children from all age group.

**Describe how the proposed sample group will be formulated:**

Data collection agency which made the data available informed that a call for representatives was sent out to targeted groups and communities in the areas of representation identified, and through the use of cascading methods.

**Indicate clearly what the involvement of the sample group will be in the research process, How their consent will be obtained, potential risks to them, Anonymity of data being collected:**

The main purpose of the sample group is to help us in collecting data which will help in the development of better healthcare devices and technologies which can be used for the betterment of the society. The result of this exercise was to develop a cutting edge technology which can be further sent to healthcare doctors for their review. Data that will be collected from customer will not have their name, address or any information that promotes racism. In cases of images being collected, all the meta-data from images will be removed to find any backpropagation way to locate the Participants. All the hardware devices use for collecting data would be properly sanitized and softwares would go through several antivirus scan so as to avoid any form of data leakage from our database. Consent of participants was taken on a paper, duly signed, by data collecting organization while collecting data

**Indicate any potential risks to people using the product how you propose to minimize these:**

The developed product will be a part of academic curriculum and unless it has cleared clinical trials, it should only be used a reference and not as a actual result. People should not use the product without getting validated from a certified health official. Results displayed by the products are meant to assist healthcare professionals and not replace them.

**Ownership of data collected**

Organization that collected data and made it open source is the full owner of data and the data has been provided to us mainly for academic research purpose.

**References**

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31.

Brown, A., Kang, J., and Gjestland, T. (2011). Towards standardization in soundscape preference assessment. Applied Acoustics, 72(6):387–392.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

Coates, A. and Ng, A. Y. (2012). Learning feature representations with K-means. Neural Networks: Tricks of the Trade.

Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. *Multiple Classifier Systems*, *34*(8), 1-17.

Daniel Steele, Dirkjan Krijnders, C. G. (2013). The sensor city initiative: Cognitive sensors for soundscape transformations. Geoinformatics for City Transformation.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer series in statistics New York.

Gaver, W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. Ecological psychology, 5(1):1–29.

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., and Plumbey, M. (2013a). Detection and classiﬁcation of acoustic scenes and events. an IEEE AASP ..., pages 1–6.

Guastavino, C. (2007). Categorization of environmental sounds. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie exp´erimentale, 61(1):54.