**MAJOR PROJECT II BACKGROUND AUDIO TAGGING SYSTEM**

**Submitted By:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **Branch** |
| Bhavya Joshi | R164216015 | CSE IOT |
| Keshav Gupta | R164216026 | CSE IOT |
| Mayur Tote | R164216036 | CSE IOT |
| Prakhar Garg | R164216043 | CSE IOT |

**Under the guidance of**

## Mr. Vidyanand Mishra

**Assistant Professor, Department of Systemics**



**SCHOOL OF COMPUTER SCIENCE UNIVERSITY OF PETROLEUM & ENERGY STUDIES**

## Dehradun – 248007 Jan2020 - May2020

**Approved By**

**(Mr. Vidyanand Mishra) (Dr. Neelu Jyoti Ahuja) Project Guide Head of Department**

**DECLARATION**

I hereby declare that this submission is my own and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other Degree or Diploma of the University or other Institute of Higher learning, except where due acknowledgement has been made in the text.

Bhavya Joshi (Enroll No. R164216015) - Keshav Gupta (Enroll No. R164216026) – Mayur Tote (Enroll No. R164216036) - Prakhar Garg (Enroll No. R164216043) -

# CERTIFICATE

This is to certify that the project titled Background Audio Tagging System submitted by Bhavya Joshi (Enroll. No. R164216015), Keshav Gupta (Enroll. No. R164216026), Mayur Tote (Enroll. No. R164216036) and Prakhar Garg (Enroll. No. R164216043) to the University of Petroleum & Energy Studies, for the award of the degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING is a bonafide record of project work carried out by them under my supervision and guidance. The content of the project, in full or parts have not been submitted to any other Institute or University for the award of any other degree or diploma.

(Date: 17 April 2020) **Mr. Vidyanand Mishra**

## (Project Guide)

**ACKNOWLEDGEMENT**

We wish to express our deep gratitude to our guide Mr. Vidyanand Mishra, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his constant support and valuable suggestions.

We sincerely thank to our Head of the Department, Dr. Neelu Jyoti Ahuja, for her great support in doing our Background Audio Tagging System at SoCS.

We would like to thank all our friends for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our parents who have shown us this world and for every support they have given us.

|  |  |  |  |
| --- | --- | --- | --- |
| Name Bhavya Joshi | Keshav Gupta | Mayur Tote | Prakhar Garg |
| Roll No. R164216015 | R164216026 | R164216036 | R164216043 |

# ABSTRACT

The following will demonstrate how to apply Deep Learning techniques to the classification of environmental sounds, specifically focusing on the identification of particular urban sounds. When given an audio sample in a computer readable format (such as a .wav file) of a few seconds duration, we want to be able to determine if it contains one of the target urban sounds with a corresponding Classification Accuracy score.

To do this, we are going to create a visual representation of each of the audio samples which will allow us to identify features for classification, using the same techniques used to classify images with high accuracy.

**Keywords**: Audio tagging, audio dataset, data collection

# TABLE OF CONTENTS

1. Introduction… 1
2. Literature Review 2
3. Problem Statement 3
4. Objective 4
5. Methodology 5
6. System Requirement 6
7. Design 7

Algorithm 7

Flowchart 8

1. Implementation 9
2. Pert Chart 10
3. Result Analysis 11
4. Conclusion 12
5. Future Scope 13
6. References… 14

**LIST OF FIGURES**

* 1. Figure 1: Overview of a single-tag tagging system 1
  2. Figure 2: Flowchart of the model… 8
  3. Figure 3: Output .csv file 9
  4. Figure 4: Pert chart… 10

**INTRODUCTION**

The sounds in our everyday environment carry a huge amount of information of the events occurring nearby. Humans are able to recognize and discern many sound events but state-of- the- art automatic processing of sounds by machines still lags far behind. Further research is needed to develop robust systems capable of recognizing a wide range of sound events in realistic audio streams.

In this project, we focus on general-purpose audio tagging using a dataset of 10 categories and almost 8k of training data. Specifically, the goal of this task is to build an audio tagging system that can categorize an audio clip as belonging to one of a set of 41 diverse categories drawn from the AudioSet Ontology (related to musical instruments, human sounds, domestic sounds, animals, etc.).

In particular, this task deals with user-generated audio clips retrieved from Free sound, which are very diverse in terms of acoustic content, recording techniques, clip duration, etc. Likewise, these audio clips sometimes feature incomplete and inconsistent user-provided metadata.

Therefore, this task addresses two main challenges of:

1. Recognizing an increased number of diverse sound events, and
2. Leveraging subsets of training data featuring annotations of varying reliability.

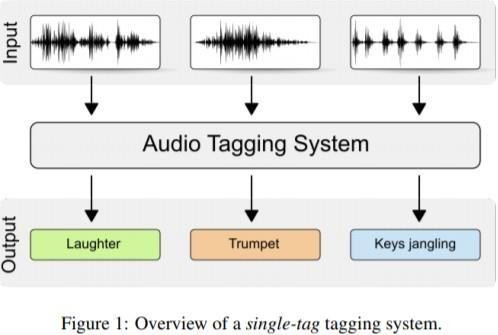
This task will provide insight towards the development of broadly-applicable sound event classifiers. Potential applications include automatic description of multimedia content, and acoustic monitoring applications.

Figure1represents the classification of audio data into its respective tags.[6]

# LITERATURE REVIEW

These sound excerpts are digital audio files in .wav format. Sound waves are digitized by sampling them at discrete intervals known as the sampling rate (typically 44.1kHz for CD quality audio meaning samples are taken 44,100 times per second).[1]Each sample is the amplitude of the wave at a particular time interval, where the bit depth determines how detailed the sample will be also known as the dynamic range of the signal (typically 16bit which means a sample can range from 65,536 amplitude values).Automatic environmental sound classification is a growing area of research with numerous real-world applications. Whilst there is a large body of research in related audio fields such as speech and music, work on the classification of environmental sounds is comparatively scarce. Likewise, observing the recent advancements in the field of image classification where convolutional neural networks are used to classify images with high accuracy and at scale, it begs the question of the applicability of these techniques in other domains, such as sound classification.[3]

# PROBLEM STATEMENT

The following will demonstrate how to apply Deep Learning techniques to the classification of environmental sounds, specifically focusing on the identification of urban sounds. When given an audio sample in a computer readable format (such as a .wav file) of a few seconds duration, we want to be able to determine if it contains one of the urban sounds.

# OBJECTIVES

The goal of this task is to predict the category for each audio clip in a test set. The task setup is a multiclass classification problem, and hence the systems to be developed in this task can be denoted as single-tag audio tagging systems.

# METHODOLOGY

We will follow the following methodology for building this model:

Data Exploration: We will do a visual inspection of the audio dataset, we can see that it is tricky to visualize the difference between some of the classes. Particularly, the waveforms for repetitive sounds thus we will do a deeper dive into each of the audio files properties.

Data Preprocessing: Librosa is a Python package for music and audio processing by Brian McFee and will allow us to load audio in our notebook as a NumPy array for analysis and manipulation.

Extract Features: To do this, we are going to create a visual representation of each of the audio samples which will allow us to identify features for classification.

Building our model: The next step will be to build and train a Deep Neural Network with these data sets and make predictions. Here we will use a Convolutional Neural Network (CNN).

MFCC: Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MF[C.[1]](https://en.wikipedia.org/wiki/Mel-frequency_cepstrum#cite_note-1) They are derived from a type of [cepstral](https://en.wikipedia.org/wiki/Cepstrum) representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the [cepstrum](https://en.wikipedia.org/wiki/Cepstrum) and the mel- frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in [audio compression](https://en.wikipedia.org/wiki/Data_compression#Audio).

# SYSTEM REQUIREMENTS (SOFTWARE/HARDWARE)

Software: Anaconda (Jupyter Notebook) Hardware: Minimum Requirements RAM: 4GB

Free Space: 8GB

# DESIGN

* Algorithm

STEP 1: Exploratory Data Analysis

* + Loading data
  + Distribution of Categories
  + Reading Audio Files
  + Audio Length

STEP 2: Building a Model using Raw Wave

* + Model Description
  + Configuration
  + DataGenerator class
  + Normalization
  + Training 1D Conv
  + Ensembling 1D Conv Predictions

STEP 3: Introduction to MFCC

* + Generating MFCC using Librosa

STEP 4: Building a Model using MFCC

* + Preparing Data
  + Normalization
  + Training 2D Conv on MFCC
  + Ensembling 2D Conv Predictions

STEP 5: Ensembling 1D Conv and 2D Conv Predictions STEP 6: Results and Conclusion

* Flowchart

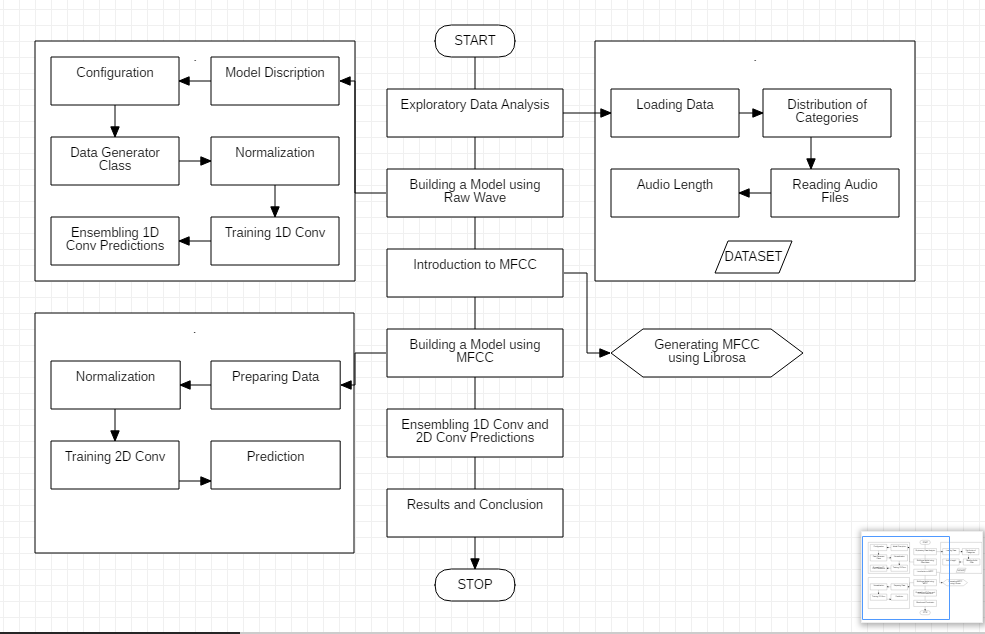
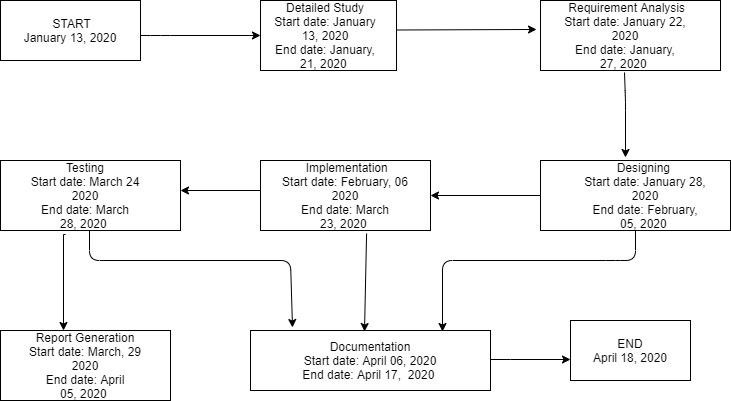


Figure 2: Flowchart of the mod

# PERT CHART



**RESULT ANALYSIS**

Our trained model obtained a Training accuracy of 93% and a Testing accuracy of 88%. The performance is very good and the model has generalized well, seeming to predict well when tested against new audio data.

The goal of this task is to predict the category for each audio clip in a test set. The task setup is a multiclass classification problem, and hence the systems to be developed in this task can be denoted as single-tag audio tagging systems

# CONCLUSION

Audio in the real world is much more of a ‘messier’ challenge, as we will need to accommodate for different background sounds, different volume levels of the target sound and the likelihood of echoes. Doing this in Real-time also poses its challenges as the model will have to perform well with low-latency, as will the MFCC calculation, all whilst synchronizing with the audio buffer thread without any delays. The dataset is currently available on the Kaggle competition page, and future updates of the dataset (including ground-truth data for the test set and extra associated Audio metadata) will be made publicly available in the Audio Datasets platform.

# FUTURE SCOPE

There is a plethora of real-world applications for this research, such as:

* Content-based multimedia indexing and retrieval
* Assisting deaf individuals in their daily activities
* Smart home use cases such as 360-degree safety and securitycapabilities
* Industrial uses such as predictive maintenance

We primarily focused on finding a technique that efficient learns strongly augmented data. We also tried to ensemble models of various low-level modulus and sampling rate, and it achieved the state- of-the-art results.

# REFERENCES

1. T. Virtanen, M. D. Plumbley, and D. Ellis, “Computational Analysis of Sound Scenes and Events. Springer”,Columbia University, NewYork 2018.
2. A. Mesaros, T. Hei’ttola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017, submitted.
3. A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 26, no. 2, pp. 379–393, 2018.
4. P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, “Chime-home: A dataset for sound source recognition in a domestic environment,” in Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2015, pp. 1–5.
5. https:[//www.intechopen.com/books/from-natural-to-artificial-intelligence-algorithms-and-](http://www.intechopen.com/books/from-natural-to-artificial-intelligence-algorithms-and-) applications/convolutional-neural-networks-for-raw-speech-recognition
6. Eduardo fonseca1\_, manoj plakal2, frederic font1, daniel p. w. ellis2,xavier favory1, jordi pons1, xavier serra1, “general-purpose tagging of freesound audio with audioset labels: task description, dataset, and baseline”, 1music technology group, UniversitatPompeuFabra, 19-20 November 2018, Surrey, UK