**Nitte Meenakshi Institute of Technology,**

Department of Computer Science and Engineering

# 18CSE751 Introduction to Machine Learning Learning Activity Proposal

**Urban Sound Classification**

**Mukesh Goit, Umang Harlalka , Abhinash Prasad Sah**

**Abstract**

This learning activity project is a simple audio classification model based on machine learning and deep learning. We address the problem of classifying the type of sound based on short audio signals and their generated spectrograms, from labeled sounds belonging to 10 different classes during model training. In order to meet this challenge, we use a model based on Convolutional Neural Network (CNN),Artificial neural network(ANN) or KNN. The audio was processed with Mel-frequency Cepstral Coefficients (MFCC) into what's commonly known as Mel spectrograms, and hence, was transformed into an image. Our final CNN model achieved 91% accuracy on the testing datasets.

# Introduction

Sounds are the part of our daily lives, ranging from the conversations we have when interacting with people, the music we listen to, and all the other environmental sounds that we hear on a daily basis such as a car driving past, the patter of rain, or any other kind of background noise. Sound classification is a constantly developing area of research and is at the heart of a lot of advanced technologies including automatic speech recognition systems, security systems, and text-to-speech applications. There are numerous applications which are continuously improving like video indexing and content based retrieval, speaker and sound identification use cases and potential security applications. Moreover, we know convolutional neural networks (CNNs) are widely used in image classification and they achieve significantly high accuracy, so we try to use this technique in a seemingly different field of audio classification, where discrete sounds happen over time. This project aims to build a deep learning powered audio classifier. The basic underlying problem to be able to manipulate audio data and build a model to classify sounds. The project aims to leverage progress achieved in the deep learning field for speaker identification and recognition problem in order to perform accurately and improve to effectively assist users. The input of the algorithm used in this project is taken from the database of the Urban Sound Classification Challenge, which are short audio samples commonly found in an urban environment like children playing, street music, a car engine etc. The samples must first be pre-processed in order to extract the MFCC features of each audio signal. the MFCC features vector is used as an input for the CNN model for classification and to generate predictions. The neural network outputs a vector with the probabilites of the sample belonging to each of the registered class. This vector is used to generate a prediction of the class of the sound, guessing for the one with the highest probability.

# Data Set

We will use the dataset from the Kaggle. The dataset is obtained from the following link –

<https://www.kaggle.com/chrisfilo/urbansound8k.>

The dataset is present wav file and will contain 8732 entries of unique values.

# Machine Learning Methods

* **ANN:** An ANN is based on a collection of connected units or nodes called [artificial neurons](https://en.wikipedia.org/wiki/Artificial_neuron" \o "Artificial neuron), which loosely model the [neurons](https://en.wikipedia.org/wiki/Neuron" \o "Neuron) in a biological brain. Each connection, like the [synapses](https://en.wikipedia.org/wiki/Synapse" \o "Synapse) in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal then processes it and can signal neurons connected to it. The "signal" at a connection is a [real number](https://en.wikipedia.org/wiki/Real_number" \o "Real number), and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called *edges*. Neurons and edges typically have a *[weight](https://en.wikipedia.org/wiki/Weighting" \o "Weighting)* that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.
* **CNN:** Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs. They have three main types of layers, which are: Convolutional layer, Pooling layer and fully connected (FC) layer. The convolutional layer is the first layer of a convolutional network. While convolutional layers can be followed by additional convolutional layers or pooling layers, the fully connected layer is the final layer. With each layer, the CNN increases in its complexity, identifying greater portions of the data. Earlier layers focus on simple features.

**Mel-frequency cepstral coefficients (MFCCs):** In [sound processing](https://en.wikipedia.org/wiki/Sound_processing" \o "Sound processing), the **mel-frequency cepstrum** (**MFC**) is a representation of the short-term [power spectrum](https://en.wikipedia.org/wiki/Power_spectrum" \o "Power spectrum) of a sound, based on a [linear cosine transform](https://en.wikipedia.org/wiki/Cosine_transform" \o "Cosine transform) of a [log power spectrum](https://en.wikipedia.org/wiki/Power_spectrum" \o "Power spectrum) on a [nonlinear](https://en.wikipedia.org/wiki/Nonlinear_system" \o "Nonlinear system) [mel scale](https://en.wikipedia.org/wiki/Mel_scale" \o "Mel scale) of frequency.

**Mel-frequency cepstral coefficients** (**MFCCs**) are coefficients that collectively make up an MFC.They are derived from a type of [cepstral](https://en.wikipedia.org/wiki/Cepstrum" \o "Cepstrum) representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the [cepstrum](https://en.wikipedia.org/wiki/Cepstrum" \o "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 spectrum. This frequency warping can allow for better representation of sound, for example, in [audio compression](https://en.wikipedia.org/wiki/Data_compression" \l "Audio" \o "Data compression).

# Assessment:

Accuracy scores of various models will be used for assessment.

# Presentation and Visualization

Accuracy scores of the ANN and CNN models will be displayed. Moreover, an audio file can be used to describe the use of the models.

# Roles

Mukesh Goit – CNN implementation.

Umang Hrlalka –ANN implementation.

Abhinash prasad Shah – pre-processing of dataset and MFCCs implementation.

# Schedule

The schedule is a table of dates and tasks that you plan to complete.

# Date Tasks to be Completed

17/01/21 Tasks completed by chosen date

18/01/22 Tasks to be completed by the final report/ presentation date

# Bibliography

* + [https://ieeexplore.ieee.org/document/9358621](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8222524/)
  + <https://www.ibm.com/cloud/learn/convolutional-neural-networks/>
  + <https://www.kaggle.com/chrisfilo/urbansound8k>
  + https://www.researchgate.net/publication/346659500\_Urban\_Sound\_Classification\_Using\_Convolutional\_Neural\_Network\_and\_Long\_Short\_Term\_Memory\_Based\_on\_Multiple\_Features