*CIS 5200 Fall 2022, Project Proposal*

# Exploration of Sound Classification with Machine Learning

**Team Members:**

- Rohit Jagga; Email: `rjagga@seas.upenn.edu`
- Shyam Mehta; Email: `smehta@seas.upenn.edu`
- Shubham Sharma; Email: `sshubh@seas.upenn.edu`

**home pod:** Sharanya

---

## 1 Motivation

All three of us are members of Penn Sargam, a musical group here on campus. Based on this shared background, we wanted to apply machine learning methods to a problem that is predicated on sound-based or music-related data, which is how we came across the UrbanSound8k dataset. This dataset contains various types of sounds recorded in NYC such as "street music", "children playing", "car horn", etc. which were collected as amplitude samples of sound excerpts. The dataset also includes the corresponding Mel spectogram representations of the sound samples, which we felt presented us a good opportunity to lean into our shared musical interests and dive deeper into sound processing/analysis techniques & their ML-related challenges.

## 2 Dataset

*Note: Citations were compiled with Bibtex. Full citations listed at end of document in the References section.*

Further description of the dataset: The UrbanSound8k dataset contains 8732 labeled sound excerpts ($\leq 4$ seconds) of urban sounds from 10 classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, and street_music. The classes are drawn from the urban sound taxonomy.

Link to the UrbanSound8k dataset [Urb]: https://urbansounddataset.weebly.com/urbansound8k.html.

## 3 Related Work

A lot of the prior work done with the UrbanSound8k dataset has examined various unsupervised clustering algorithms, as well as different configurations of neural network architectures. We also found that there has been some extensive work done in sound processing with this dataset, such as the SpeechBrain project. SpeechBrain is an open-source all-in-one speech toolkit that is designed to make the research and development of neural speech processing technologies easier & more efficient [Rav+21].

Research organizations such as HuggingFace have in the past done development on top of SpeechBrain for sound-intensive datasets like UrbanSound8k [Fac]. Currently, Hugging Face has released a system based on SpeechBrain that is composed of a ECAPA model coupled with statistical pooling, pretrained using the UrbanSound8k dataset. A classifier, trained with Categorical Cross-Entropy Loss, is applied on top of that.

# 4 Problem Formulation

At its core, the problem we are trying to solve in the data is a classification problem: the audio samples & Mel spectograms given in the dataset are labelled according to which sound category they fall into, with the following distribution of classes across the entire dataset:

| | |
|---|---|
| air_conditioner | 12.6% |
| car_horn | 3.54% |
| children_playing | 12.53% |
| dog_bark | 9.42% |
| drilling | 10.93% |
| engine_idling | 12.98% |
| gun_shot | 1.49% |
| jackhammer | 11.85% |
| siren | 12.03% |
| street_music | 12.61% |

Approaching the overall problem from the perspective of classification, we can test the viability of several machine learning methods to see which models are best able to classify input sound samples into their correct sound categories from the 10 types.

# 5 Methods

Before evaluating different machine learning methods, we intend to perform a Principal Components Analysis (PCA) on the dataset's input features, as a way to perform dimensionality reduction. We will also further separate the train dataset into several folds for training & validation, enabling us to perform k-fold cross-validation so as to ensure that our classifiers/models are as robust as possible with regards to performance on the test set. Then, we will compare the performance of the following ML methods listed below:

- Baseline model
  - Simple logistic regression classifier
- Advanced models
  - Logistic regression classifier (with regularization & hyperparameter tuning)
  - k-Nearest Neighbors (kNN)
  - Support Vector Machines (SVM)
  - Random Forests classifier (or Gradient Boosted)
  - Convolutional Neural Network (CNN)
- State-of-the-art models from external sources (will use published packages online to compare their performance against our baseline & advanced models)
  - Hugging Face's speechbrain ECAPA model (ECAPA-TDNN architecture that incorporates both convolutional & Time-Delay NN residual blocks)

# 6 Evaluation

We intend to evaluate the performance of our chosen ML models by comparing their accuracy & error on the test set. Here are some additional ways in which we hope to build upon the loss function and accuracy metrics used, to make our evaluation of the models more extensive:

- Accounting for the class imbalance present in the dataset by specifically using a **class-weighted loss function**, instead of the standard Cross-Entropy loss or the Log loss used in logistic regression. We could also look at loss functions such as hinge loss, which would be more relevant for the SVM model.
- Accounting for the class imbalance by **downsampling from classes overrepresented** in the dataset and **upsampling from classes underrepresented** in the dataset.
- Comparing each of the models on **different performance measures** such as recall, precision, F1 score, etc.

# 7 Project Plan

## Week 11: 11/14

- Going through the dataset to perform data cleaning, preprocessing, and exploratory data analysis (EDA).
- Do further background research on sound processing techniques, to help us better understand how best to analyze the dataset's Mel spectogram sound representations.

## Week 12: 11/21

- All: Finish exploratory data analysis, perform PCA and setup procedure for k-fold cross-validation.
- Construct the model pipeline and complete modeling portion for the baseline model (logistic classifier) and 2 of the advanced models (logistic classifier w/ improvements, and kNN). 1 member sets up model pipeline & basic model, other two members implement the 2 advanced models.

## Week 13: 11/28

- All: Complete the modeling portion for the rest of the advanced models and compare with external state-of-the-art model's performance (1 team member works on 1-2 models each).
- All: Test different loss functions, address class imbalance, test out difference accuracy measures (i.e. precision, recall, etc.)
- All: Start writing paper/presentation (each member starts writeup for their portion).

## Week 14: 12/5

- All: Complete project writeup.
- All: Presentation in pod.

# References


[Rav+21] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv:2106.04624. 2021. arXiv: 2106.04624 [eess.AS].

[Fac] Hugging Face. *Speechbrain/urbansound8k$_e$capaÂhuggingface*. URL: https://huggingface.co/speechbrain/urbansound8k_ecapa.

[Urb] UrbanSound8k. *UrbanSound8K*. URL: https://urbansounddataset.weebly.com/urbansound8k.html.