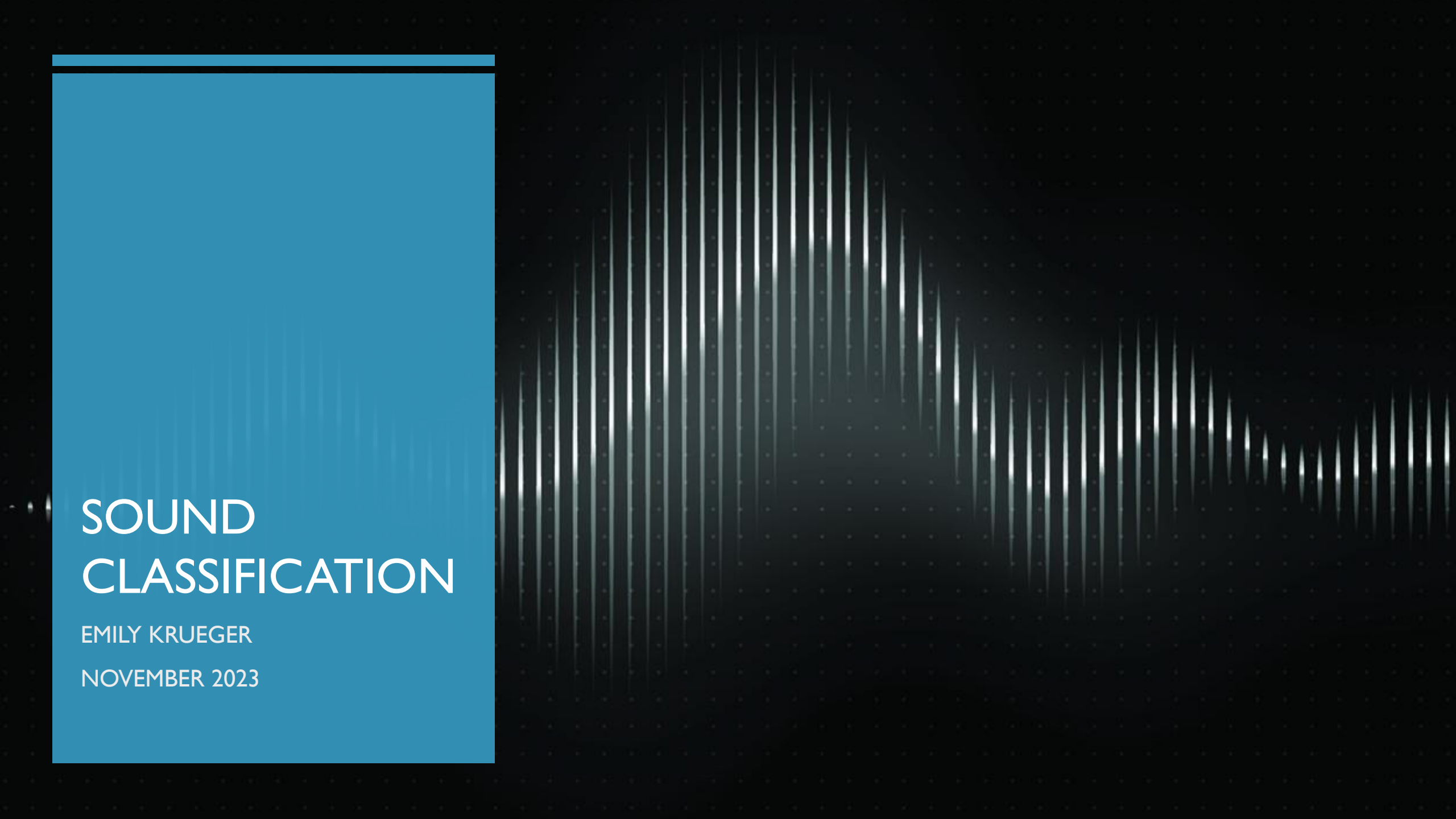
SOUND CLASSIFICATION

EMILY KRUEGER

NOVEMBER 2023



BUSINESS OVERVIEW

- **Krueger Consulting** has been contracted by a top streaming service to develop a sound classification model that can accurately distinguish between different sound classes
  - The platform hosts millions of music and podcasts
  - On a daily basis, the platform's users upload thousands of audio files
  - The company is in need of a model that can distinguish between the following classes:

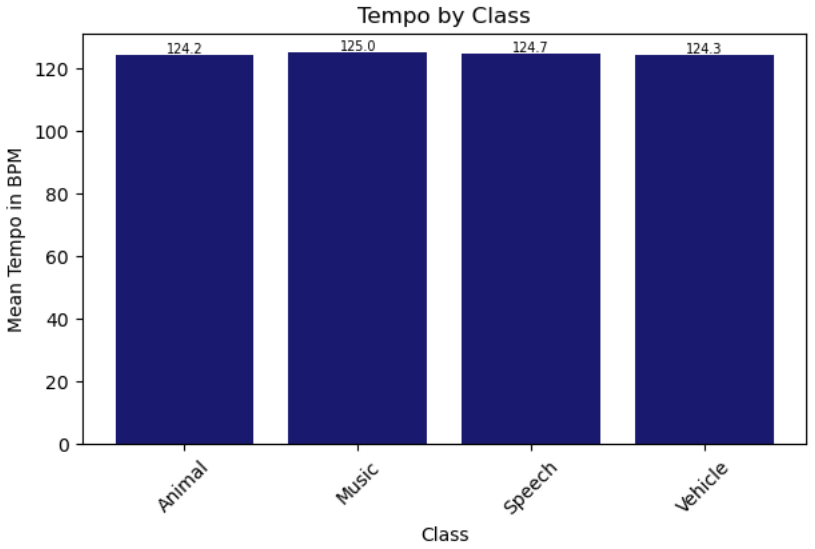
    1. Music
    2. Speech
    3. Animal
    4. Vehicle

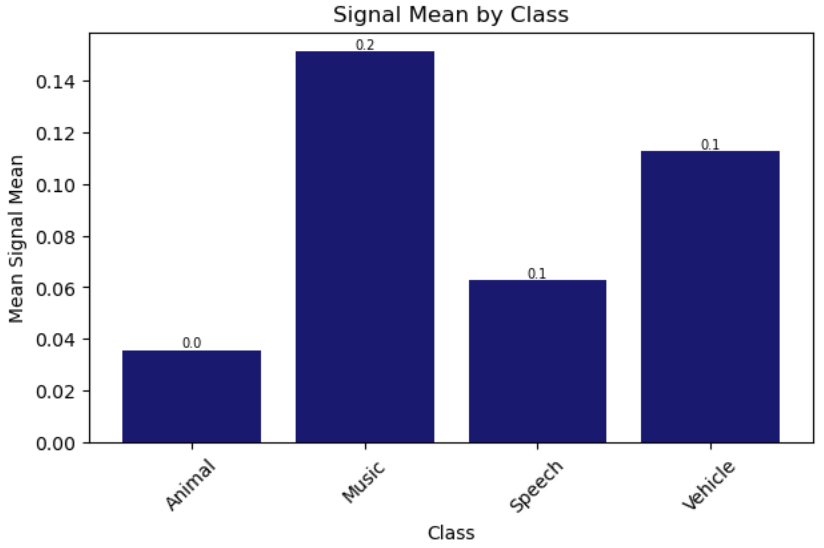
  - Our client will use this model to label and segment out all music and podcasts that users attempt to upload and discard any other sound classes

DATA OVERVIEW

- **Source:**
  - AudioSet, a dataset of approximately 2.1 million human annotated, ten second YouTube clips
    - The dataset was compiled and published by the Sound Understanding team in the Machine Perception Research group at Google
    - A subset of the larger dataset was downloaded, focusing specifically on data points labelled Music, Speech, Animal, or Vehicle
  - Resulting dataset consists of 11,513 audio files and is fairly balanced, with each class representing anywhere from 22% to 27% of the dataset
- **Strategies:**
  1. Extract numerical features from audio files to build various traditional machine learning models
  2. Convert audio files to mel spectrograms to build a convolutional neural network

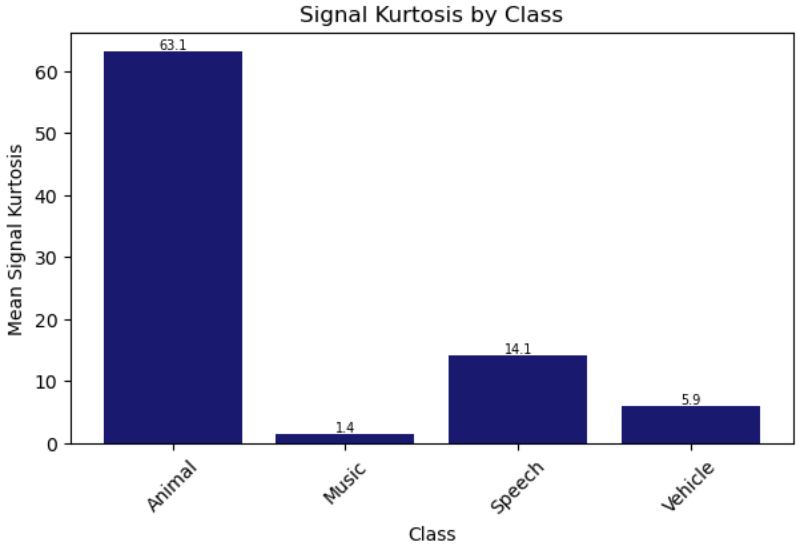
NUMERICAL FEATURES EDA

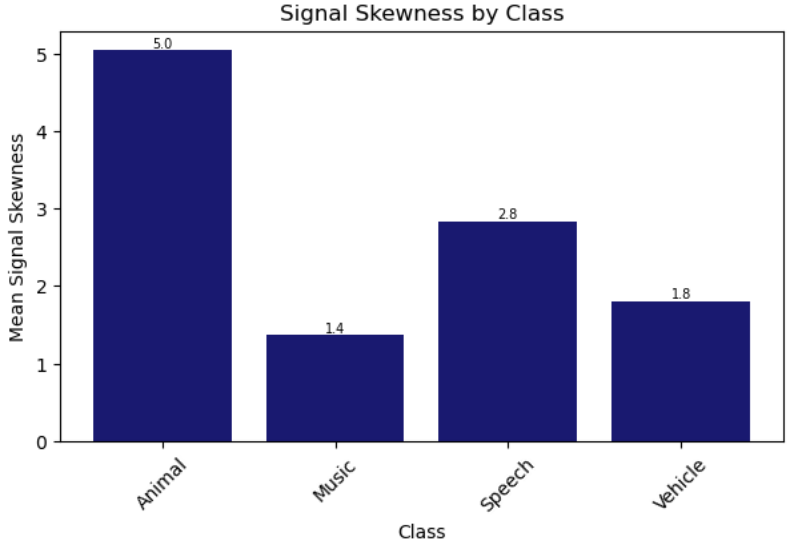




Tempo is a rhythmic feature measured in beats per minute

Signal Mean is the mean amplitude of the audio signal

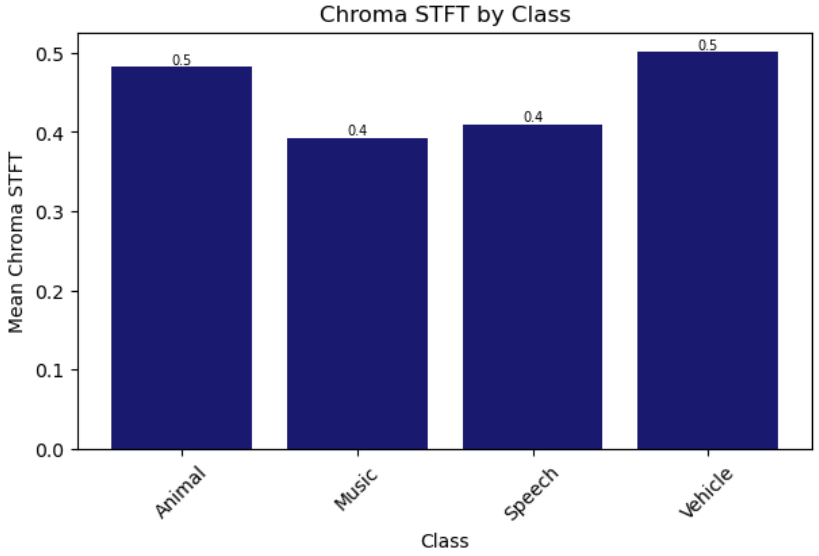


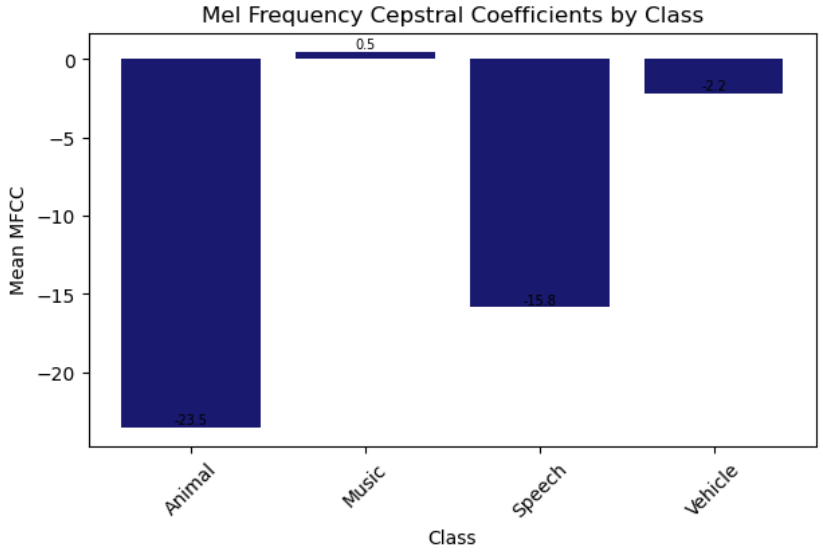


Skewness is a measure of the asymmetry of time series data

Kurtosis is a measure of the heaviness of the tails in comparison to a normal distribution

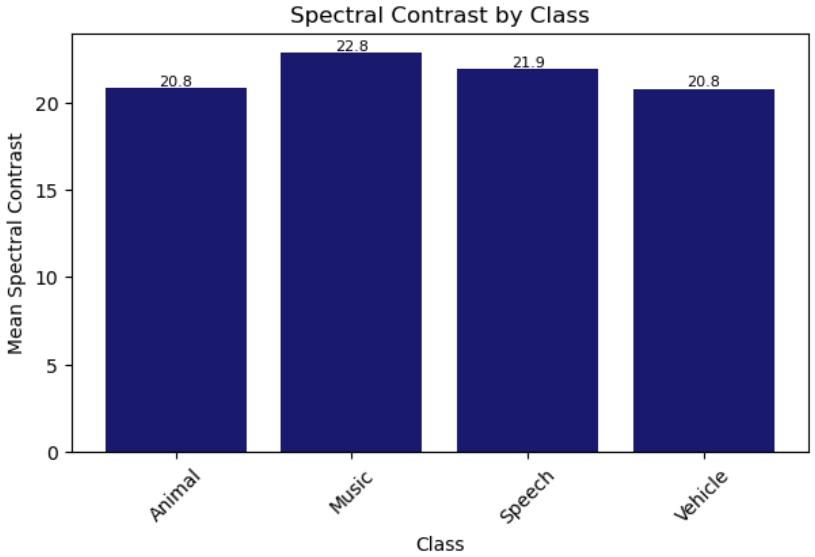
NUMERICAL FEATURES EDA

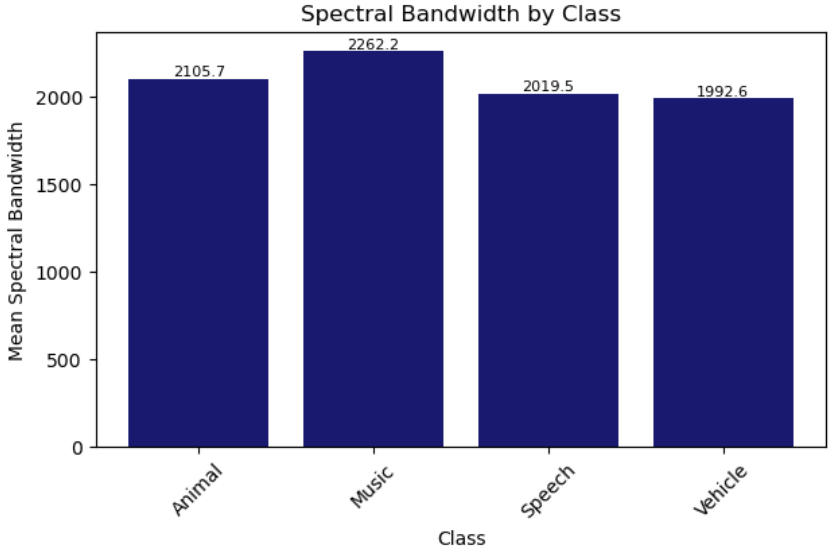




Chromagrams represent energy distribution of pitch classes

MFCCs are coefficients that represent the short-term power spectrum of the audio signal, thus capturing important spectral characteristics

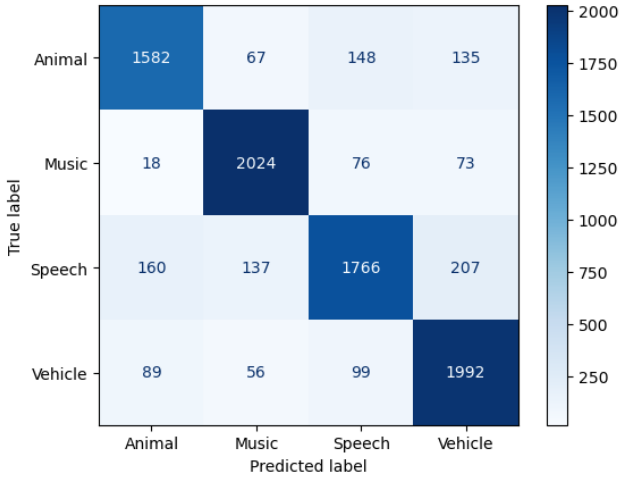




Spectral content of an audio signal refers to the distribution of energy across different frequencies

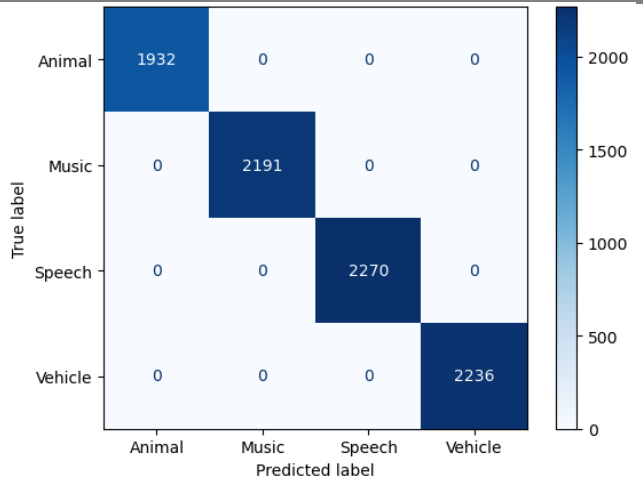
TRADITIONAL MACHINE LEARNING MODELS

Random Forest



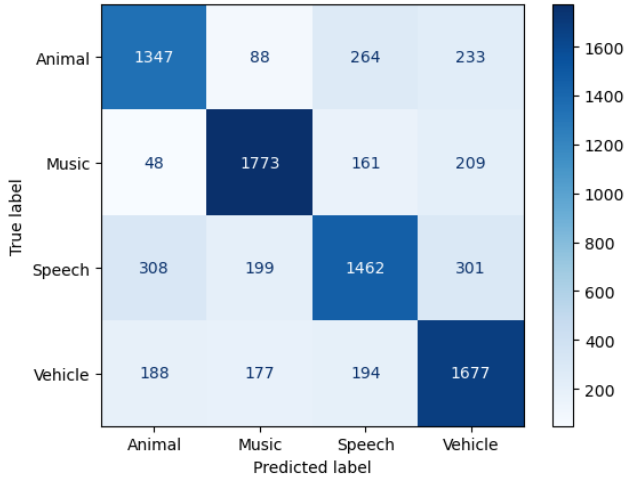
Accuracy:
Train: 88%
Test: 71%

AdaBoost Classifier



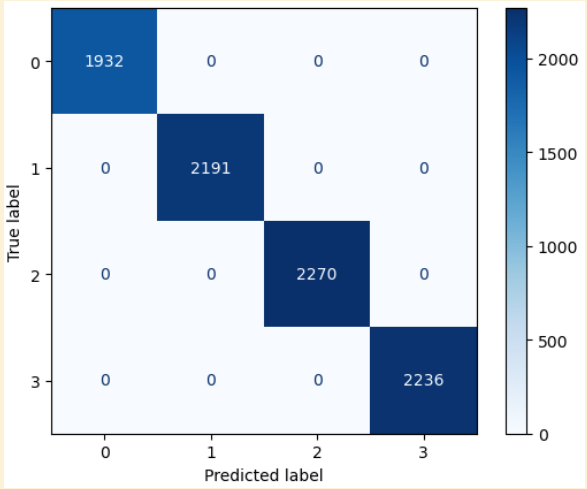
Accuracy:
Train: 100%
Test: 74%

Logistic Regression



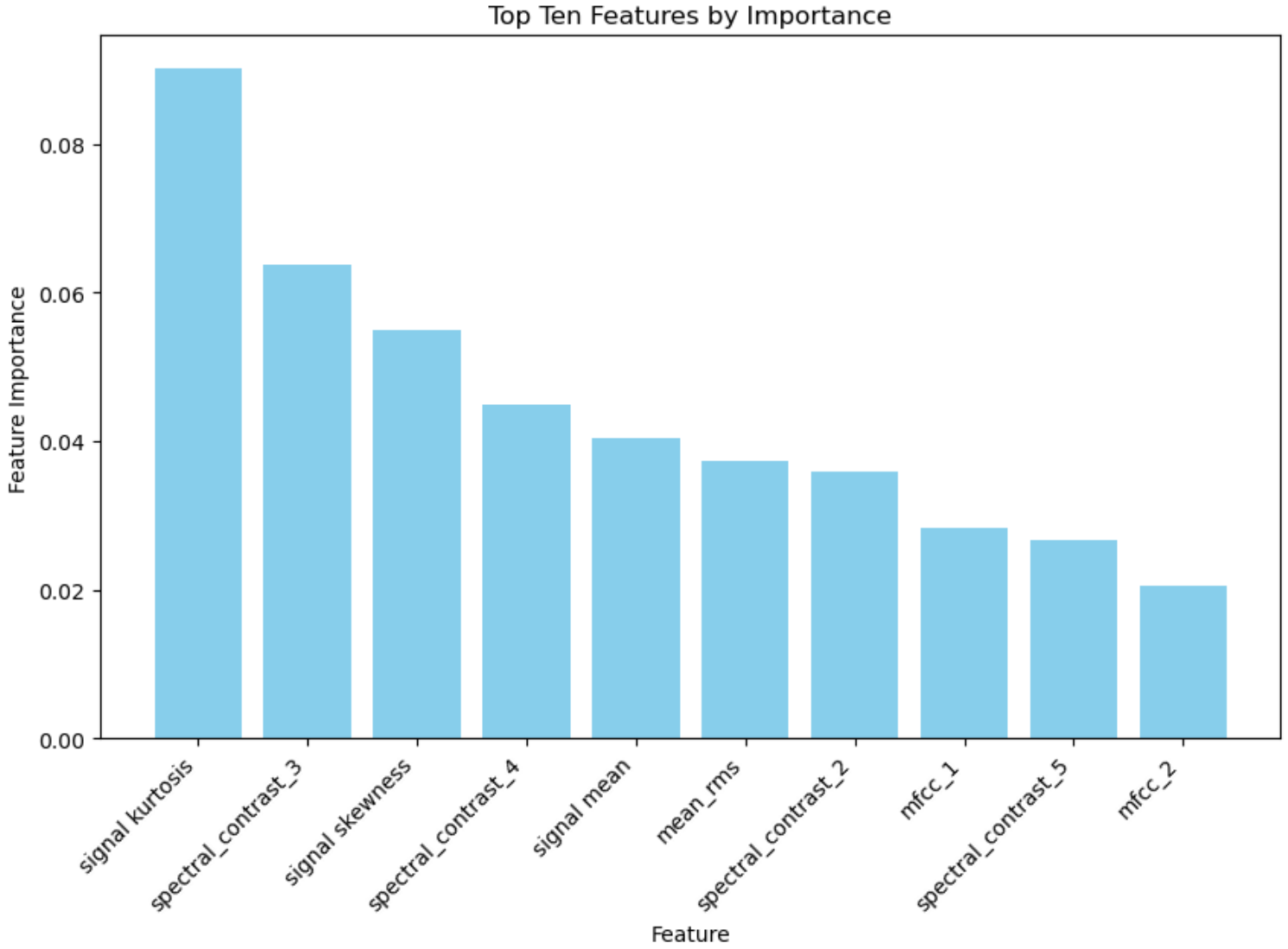
Accuracy:
Train: 73%
Test: 72%

XGBoost Classifier

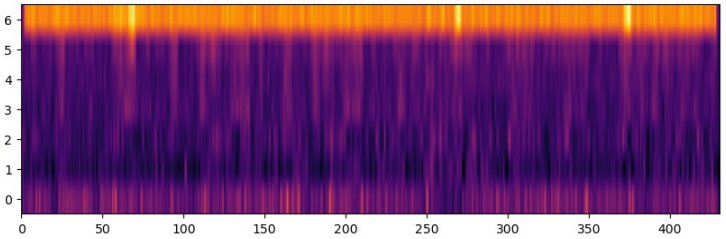


Accuracy:
Train: 100%
Test: 76%

FEATURE IMPORTANCE - XGBOOST

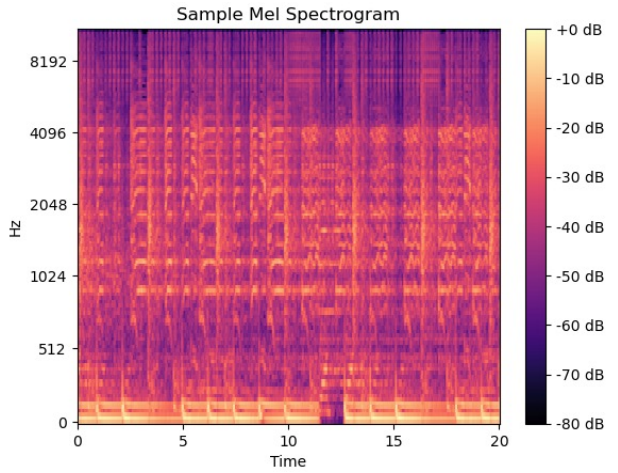


- Signal kurtosis is the most meaningful feature in determining sound classification
- Spectral contrast across four different frequency bands is also significant
- Example of visualized spectral contrast with six bands:

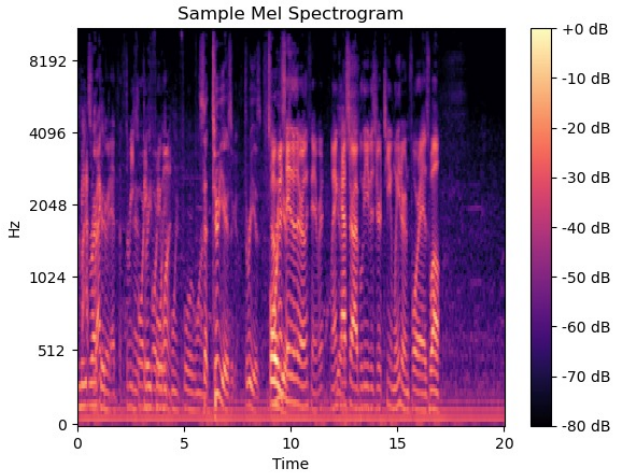


MEL SPECTROGRAMS

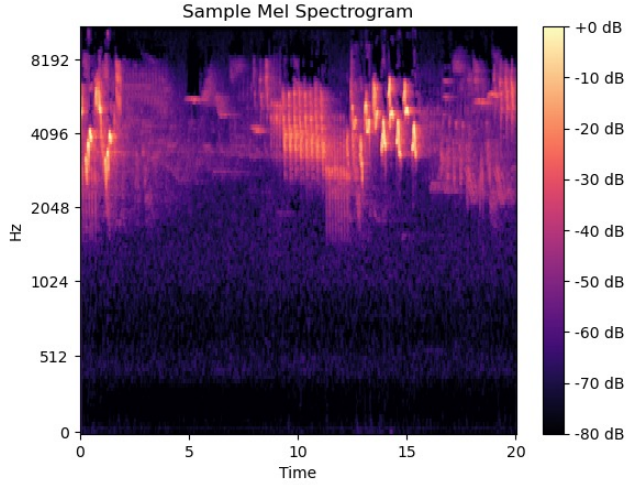
Music (EDM)



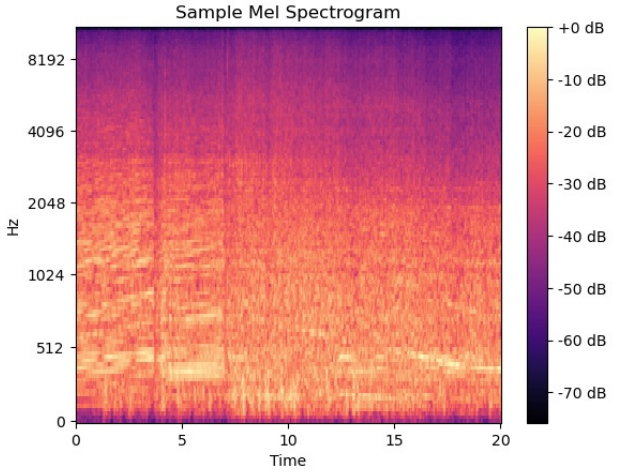
Speech (sports commentary)



Animal (Birds Chirping)



Vehicle (Engine Revving)



A Mel-Spectrogram is a visual representation of audio with time on the x-axis, frequency on the y-axis, and color representing amplitude in decibels

CONVOLUTIONAL NEURAL NETWORK (CNN)

Layer (type)	Output Shape	Param #
conv2d_9 (Conv2D)	(None, 126, 126, 32)	896
batch_normalization_9 (BatchNormalization)	(None, 126, 126, 32)	128
re_lu_9 (ReLU)	(None, 126, 126, 32)	0
max_pooling2d_9 (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_10 (Conv2D)	(None, 61, 61, 64)	18496
batch_normalization_10 (BatchNormalization)	(None, 61, 61, 64)	256
re_lu_10 (ReLU)	(None, 61, 61, 64)	0
max_pooling2d_10 (MaxPooling2D)	(None, 30, 30, 64)	0
conv2d_11 (Conv2D)	(None, 28, 28, 64)	36928
batch_normalization_11 (BatchNormalization)	(None, 28, 28, 64)	256
re_lu_11 (ReLU)	(None, 28, 28, 64)	0
max_pooling2d_11 (MaxPooling2D)	(None, 14, 14, 64)	0
global_average_pooling2d_3 (GlobalAveragePooling2D)	(None, 64)	0
dense_3 (Dense)	(None, 4)	260

Total params: 57220 (223.52 KB)
Trainable params: 56900 (222.27 KB)
Non-trainable params: 320 (1.25 KB)

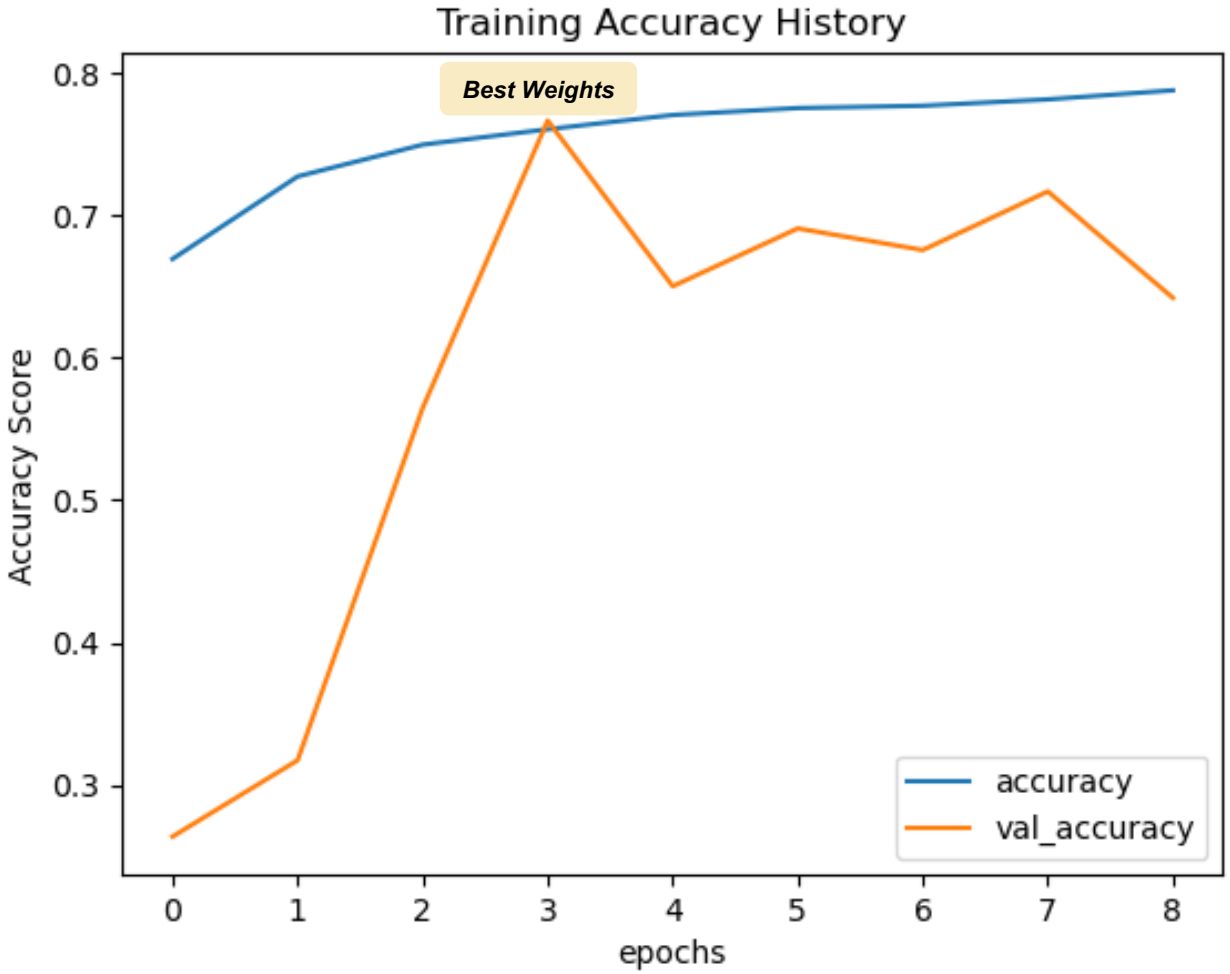
Input Layer

Layer 2

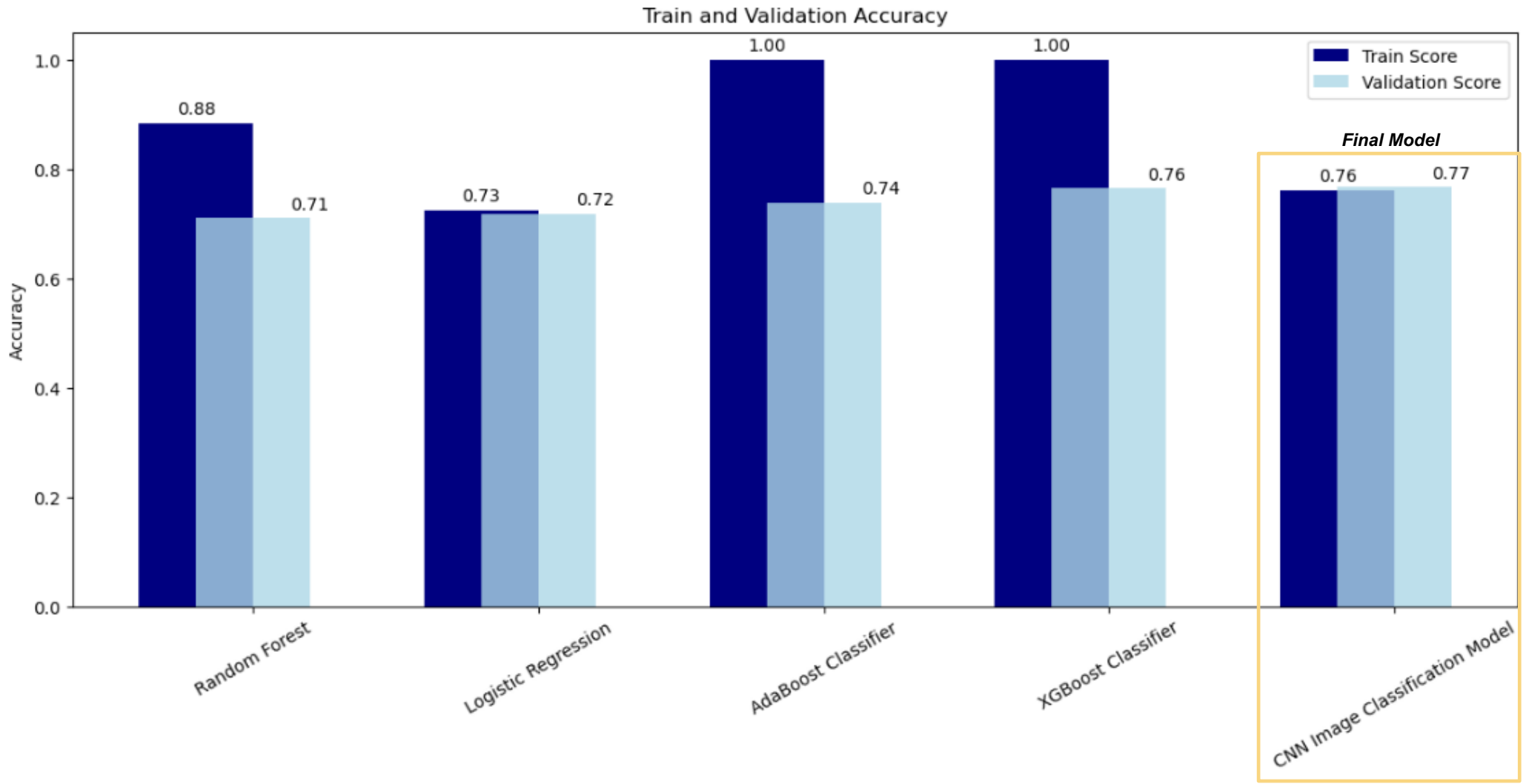
Layer 3

Final Pooling Layer

Output Layer

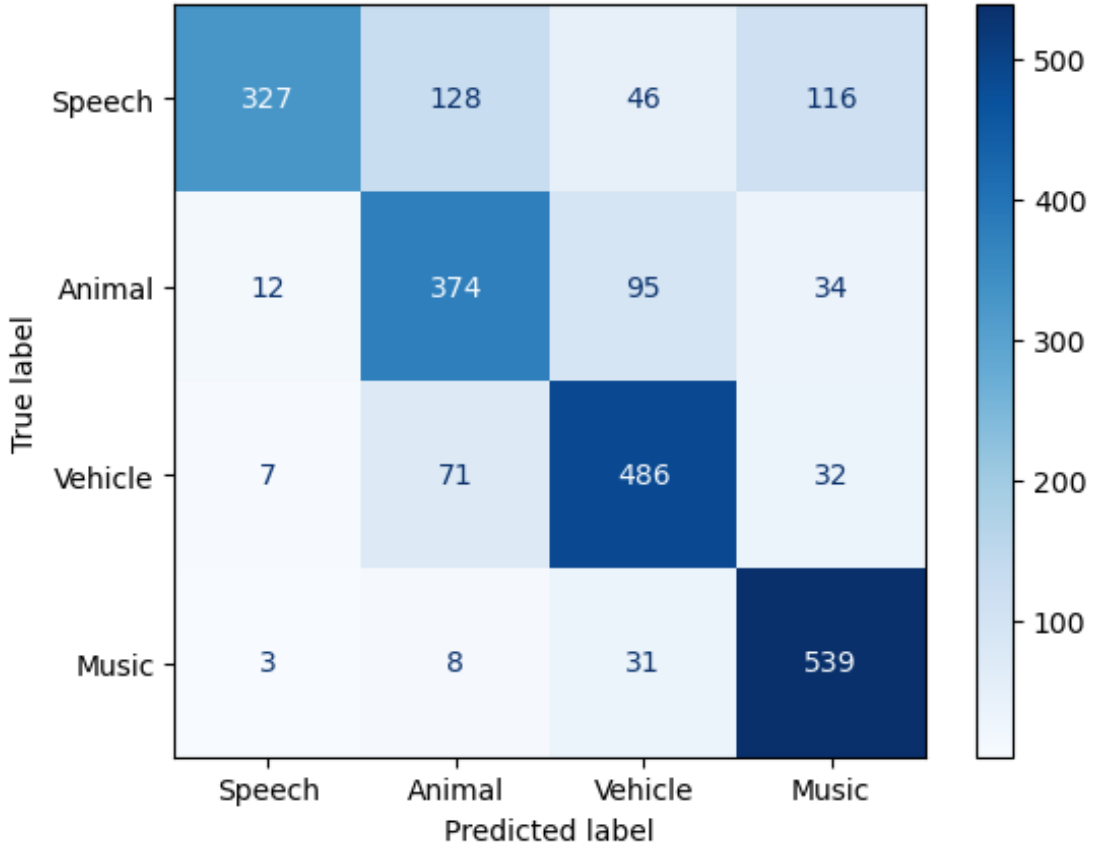


MODEL COMPARISON



CNN EVALUATION ON UNSEEN DATA

Our model was tested on unseen data obtained from the streaming platform that contracted Krueger Consulting and yielded the following results:



Accuracy: 75%

Precision: *77%*

Recall: *75%*

F1: *74%*

CONCLUSION

Recommendations:

- We recommend our client use this model to label music and podcasts as users upload audio to their platform and block/discard any animal or vehicle audio that users attempt to upload
- One could also use this model as a starting point to develop a virtual assistant that can both recognize and interpret speech and music, provide song information, or respond to requests

Additional Considerations/Next Steps:

1. Train models on a larger dataset in hope of improving accuracy score
2. Add additional sound classes to increase number of use cases

An application that leverages our final model can be found here:

https://sound-classifier-app.streamlit.app/

BIOGRAPHY

Emily Krueger

Email: ekrueger1217@gmail.com

M: 732-403-4566

Github: https://github.com/ekrueger1217

LinkedIn: https://www.linkedin.com/in/emily-krueger-058513103/