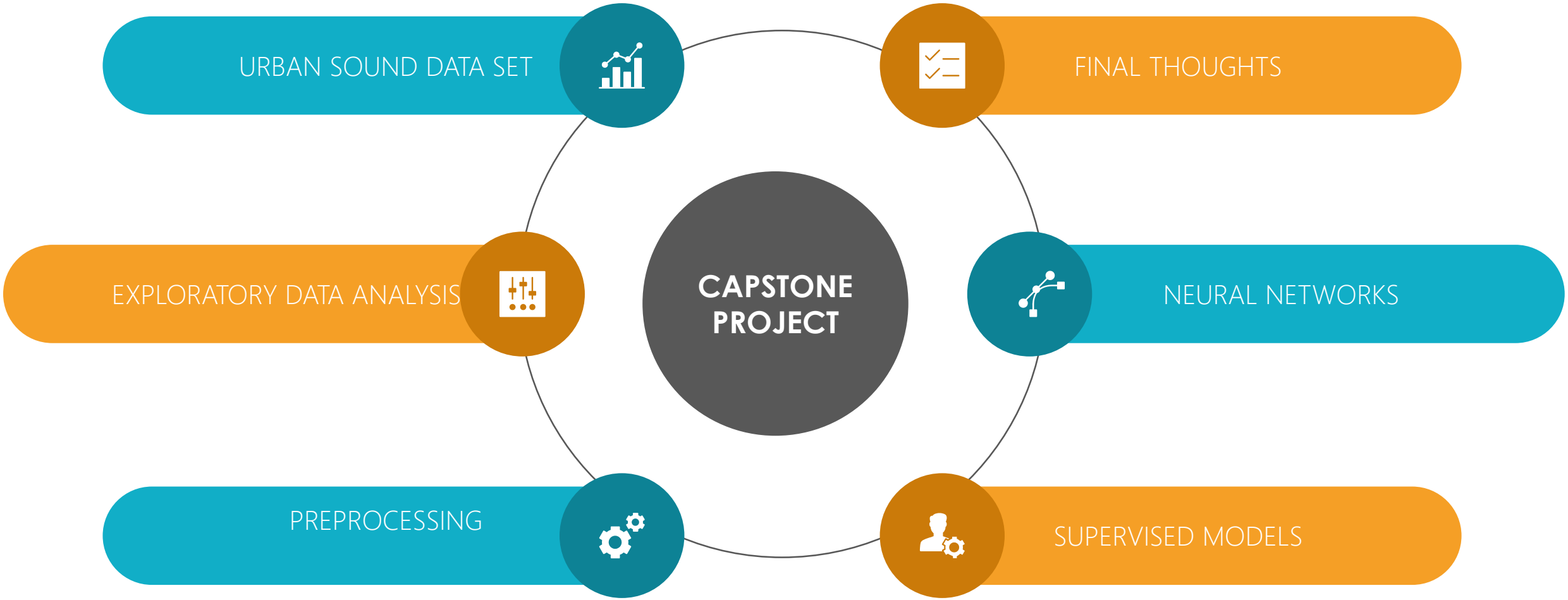
Urban Sound Classification

Final Capstone Project

URBAN SOUND CLASSIFICATION



PROJECT DESIGN

INTRODUCTION

- Sonic event classification is a field of growing research.
- Most of these researches focuses on music or speech recognition.
- Scarce works on environment sounds.
- Very few databases for labeled environment audio data.

EDA

- Urban Sound 8k Dataset contains 8723 audio excerpts in .wav formats from 10 classes of different sound sources
- We identified audio files with duration < 1 sec.
- Great variation across audio samples.

PREPROCESS

- Our biggest challenge is feature extraction.
- Audio data cannot be expressed in vector forms like other type of data such as images and texts.
- Applied various feature extraction techniques.

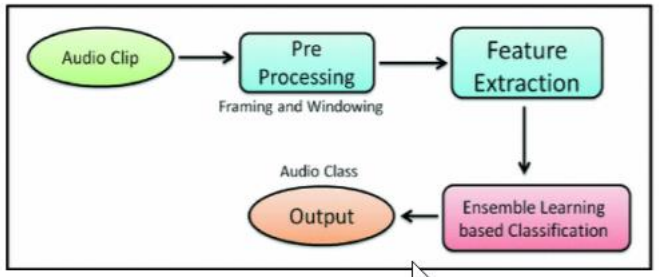
MODEL ANALYSIS

- Analyze Supervised models.
  - Random Forest
  - SVM
- Analyze Neural Networks using TensorFlow and Keras
  - Deep NN
  - CNN
- Perform Hyperparameter Optimization.

MODEL EVALUATION

- Evaluate the performance of machine learning algorithms.
- Choose the best model based on performance metrics:
  - Training Accuracy
  - Test Accuracy

IMPLEMENTATION APPROACH



Source: Audio Files Process

- 8732 Audio Files
- 10 Classes

EDA

- MFCC is a 20-dimensional features, widely used in automatic speech and speaker recognition.
- VGGish is 128-dimensional features, a pre-trained convolutional neural network.

FEATURE EXTRACTION

AUDIO DATA SET

MODEL EVALUATION

SUPERVISED MACHINE LEARNING

- Random Forest
- Support Vector Machine

NEURAL NETWORKS

- Deep NN
- Convolutional NN

PERFORMANCE METRICS

- Classification Report
- Confusion Matrix
- Accuracy Scores

IMPLEMENTATION APPROACH

PROBLEM STATEMENT

Classify the audio files in urban setting and measure the performances of various models.

a. What feature extraction techniques should be used for optimal results?
b. How do the machine learning models compare against the neural network learning models?
c. Which model performed the best?

APPROACH

a. First, perform exploratory data analysis on the audio files to quickly assess audio patterns.
b. Use feature extraction techniques for audio feature generation and embedding post processing.
c. Apply various machine learning based classification techniques to train the model to classify the audio file.
d. Evaluate and choose the best performer by measuring the effectiveness of different models.

FEATURE EXTRACTION

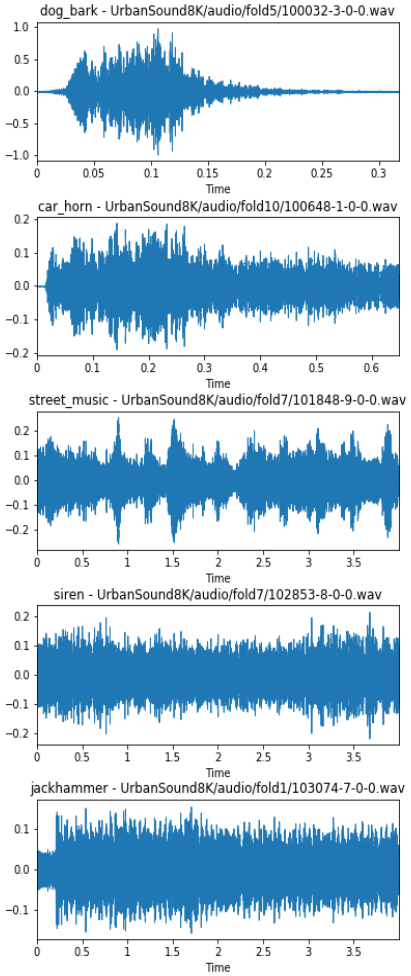
Feature extraction is the most important part for designing a machine learning model.

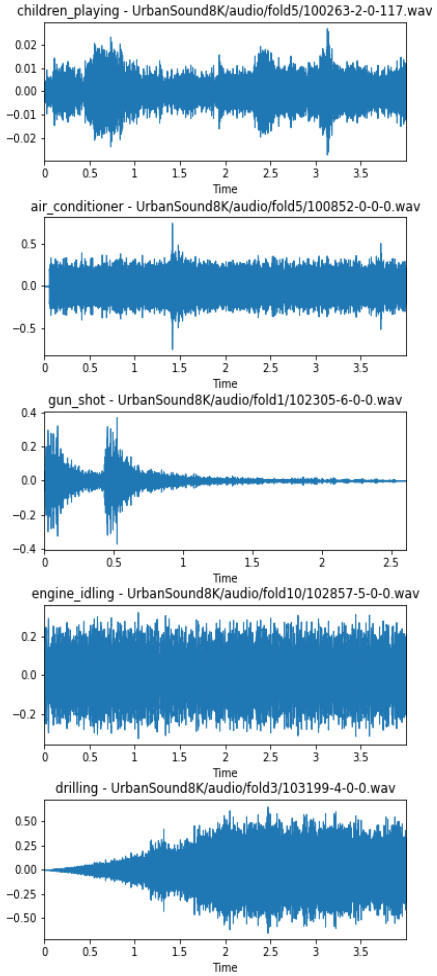
1. Mel-frequency cepstral coefficients (MFCC)
   - This method is available in Librosa Library. · It can extract 20-dimensional features from an audio file.
   - And is widely used in automatic speech and speaker recognition.
2. Visual Geometry Group (VGG, also Known as VGGish)
   - This method is available in the Audioset Library.
   - It can extract 128-dimensional features from an audio file.
   - A pre-trained convolutional neural network.

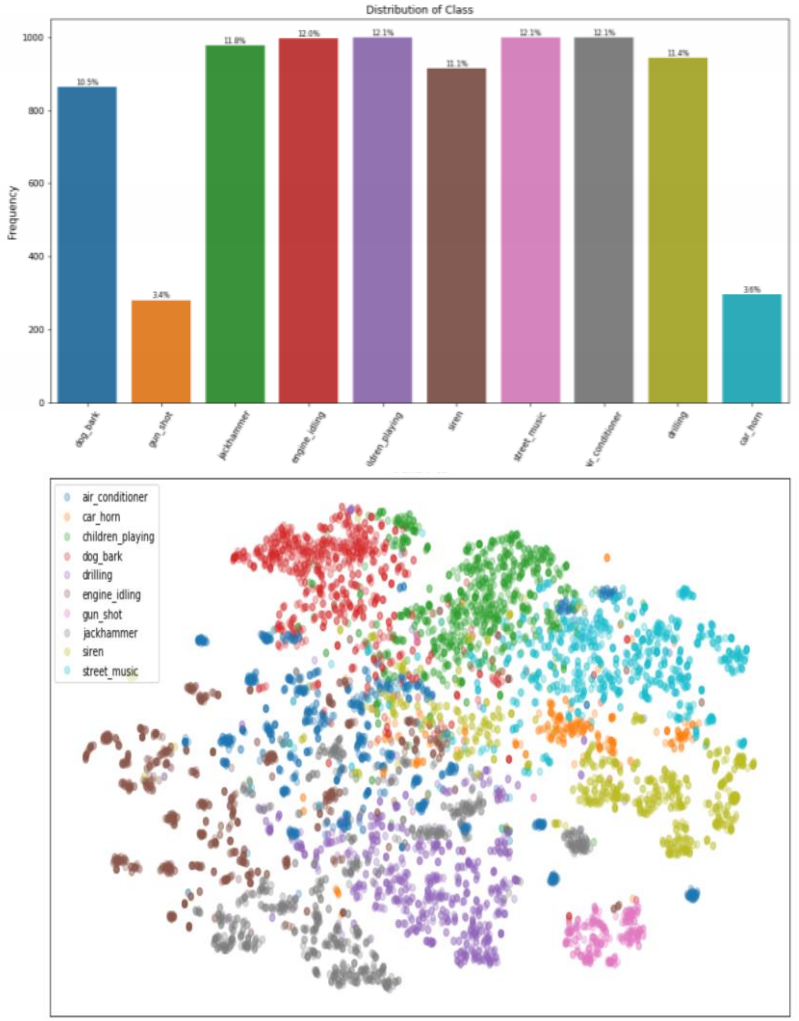
MODEL EVALUATION

1. Split the dataset into 80% train set and 20% test set.
2. Optimize and train the model with best parameters
3. Evaluate the model on test data set using performance metrics and confusion matrix.

VISUALIZATIONS







- Librosa Library was used to plot the wave plots of the audio files.
- Gunshot and dog bark had some distinct wave lengths.
- Distribution of classes shows that the gunshot and car horn had under 300 samples compared to others.
- T-SNE plot clustering plot clearly shows Gunshot has distinct cluster whereas there are some overlapping in other clusters.

HIGHLIGHTS

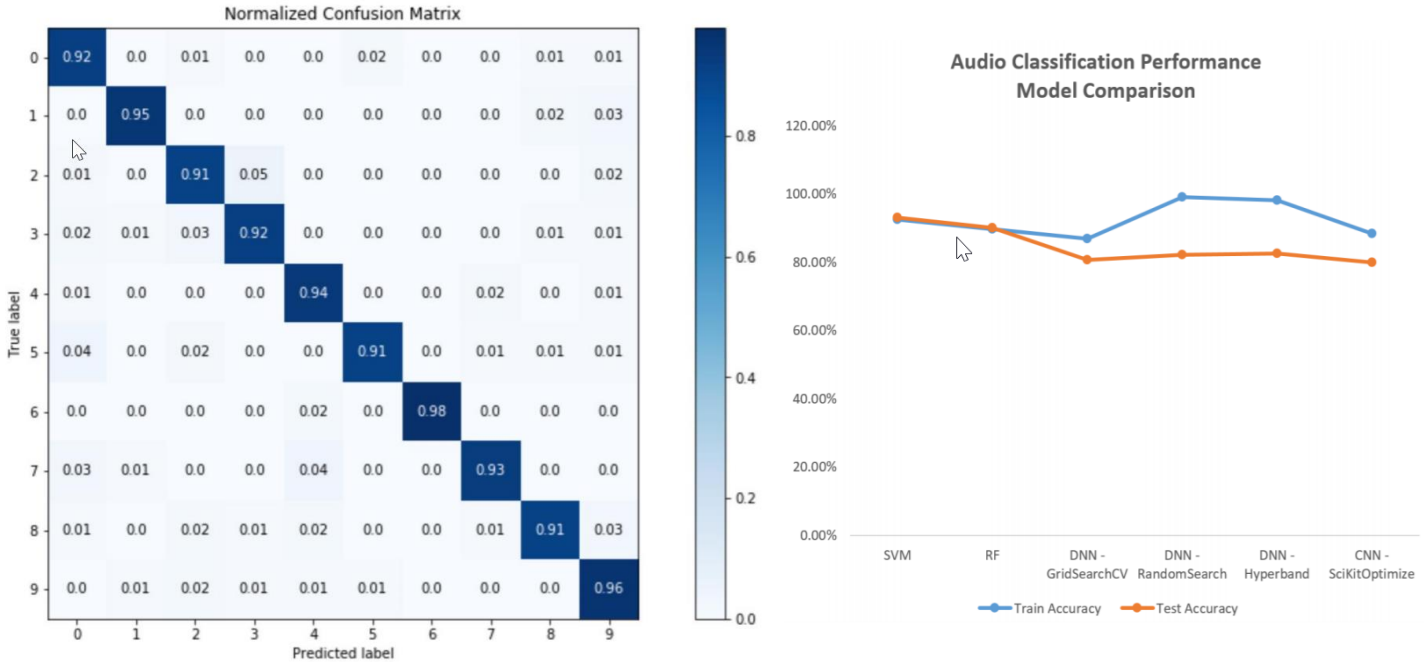


Table - Model Comparison

Model	Hyperparameter Optimization	Training Accuracy	Test Accuracy	Issues
Support Vector Machine	GridSearchCV	92.46%	93.00%	-
Random Forest	GridSearchCV	89.62%	90.00%	-
Deep Neural Network	GridSearchCV	86.90%	80.60%	Overfitting
Deep Neural Network	Keras Tuner - RandomSearch	99.00%	82.20%	Overfitting
Deep Neural Network	Keras Tuner - Hyperband	98.15%	82.50%	Overfitting
Convolutional Neural Network	Sci-Kit Optimize	88.36%	79.91%	Overfitting

- We trained 4 different models with hyper parameter optimization.
- Support Vector Machine SVM model performed better than all others with training and test accuracy of 93%.

- Both DNN and CNN seems to have overfitting issues.

- Car horn and gun shot have less than 300 samples compare to other classes, which have around 1000 samples each.

- Having the lowest number of samples, gunshot is still managed to have the highest proportion for true positive value. However, the car horn is often misclassified as the street music.

- It is difficult to differentiate between jackhammer and drilling, however it is easy to discern between dog bark and drilling.

FINAL THOUGHTS

PRACTICAL USES

The automatic classification of audio events in an urban setting has a variety of applications. Some of them are listed below.

- Audio Event Detection
- Home security or Audio Surveillance
- Assisted living, elder or infant care
- Accident and crime surveillance

Manually monitoring of urban sounds either in close proximity or remotely through a monitoring device, not only demands attention but also requires the person to be within hearing distance.

This is not always possible, and is where audio event detection, or sound recognition, solves real problems.

It will automatically alerts an application if a specific sound is detected, so that a human may take the appropriate action.

FUTURE CONSIDERATIONS

This capstone project focuses on the various machine learning techniques to model the data to give us predictive power to classify the sonic events accurately.

- Improving this model to optimize prediction of the audio classification includes supervised machine learning models such as Random Forest, and Support Vector Machines as well as neural network models such as Deep Neural Networks and Convolutional Neural Networks using Tensor Flow and Keras.

- We can potentially improve the quality of life of city dwellers by providing a data-driven understanding of urban sound and noise patterns, partly enabled by the move towards “smart cities” equipped with multimedia sensor networks.

Thank You