

Abstract

In the modern age of autonomous cars, sight is only one sense out of the five, that is being used to analyze the environment. In this research, we'd like to investigate the classification of sounds, as it can assist us in situations, where our sight is obstructed. Working off of UrbanSound8K to detect 10 sound classes, we go on to detail the effectiveness of different models for this task. After our comparisons, we find that the CNN does best. The best accuracy achieved can be up to 65%.

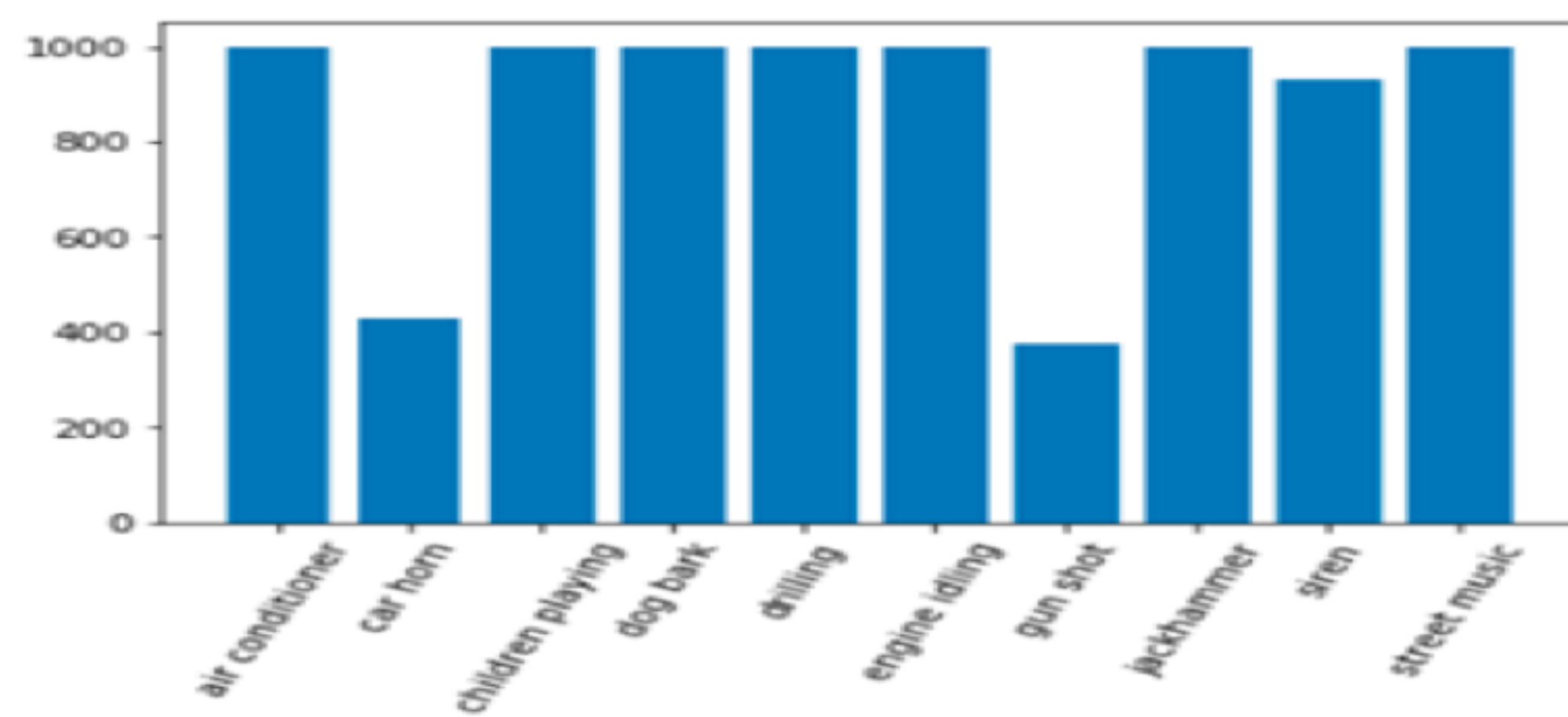
Introduction

- Worked on UrbanSound8K dataset because it is not just based off of its auditory scene type, but on the sources of its sound
- Feature extraction techniques consist of Mel Frequencies Cepstral Coefficient (MFCC), spectrograms, spectral contrasts, and tonal centroid features
- Applied general classifiers: K-Nearest Neighbor, Deep Neural Network and Convolutional Neural Network



Example of urban environment where data is collected

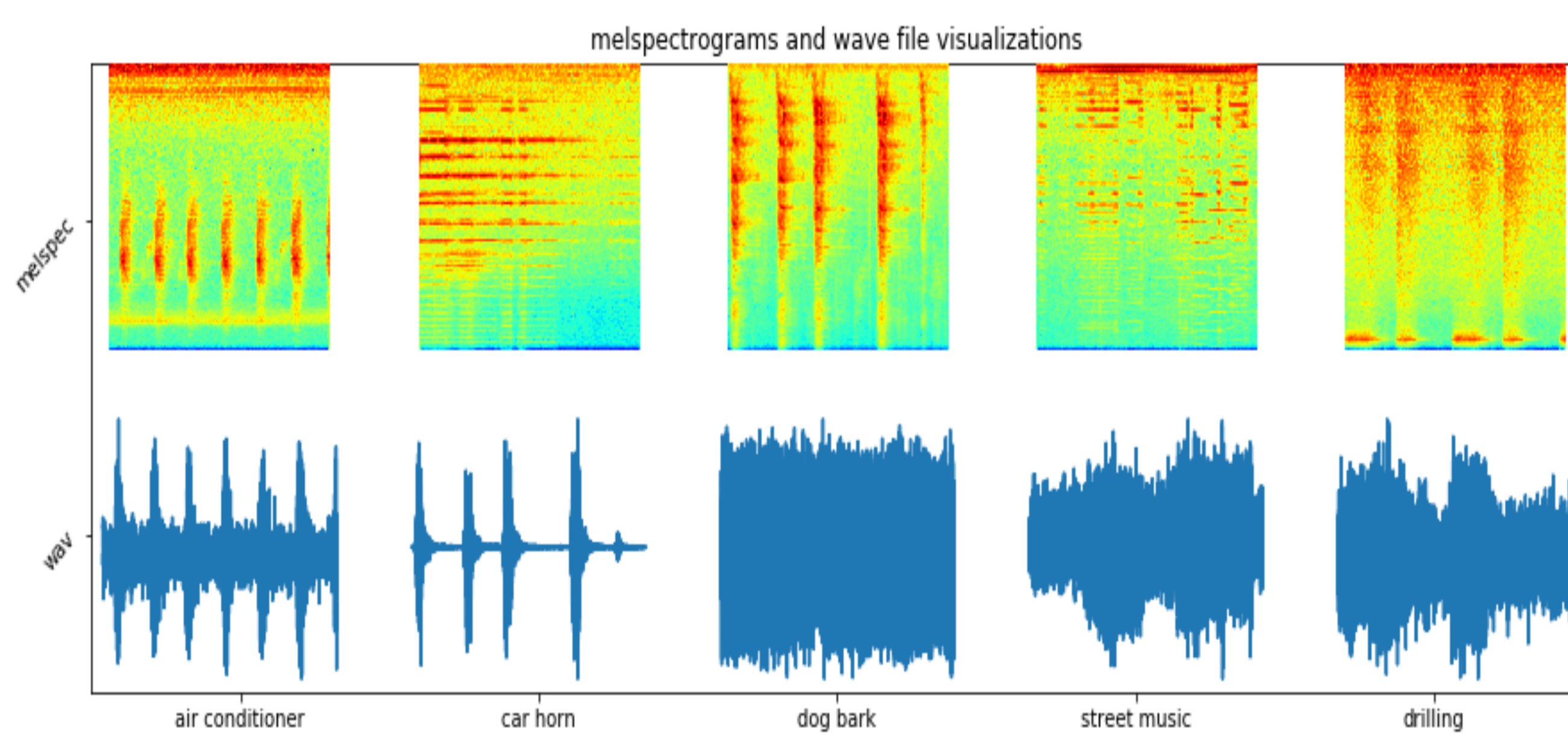
Data



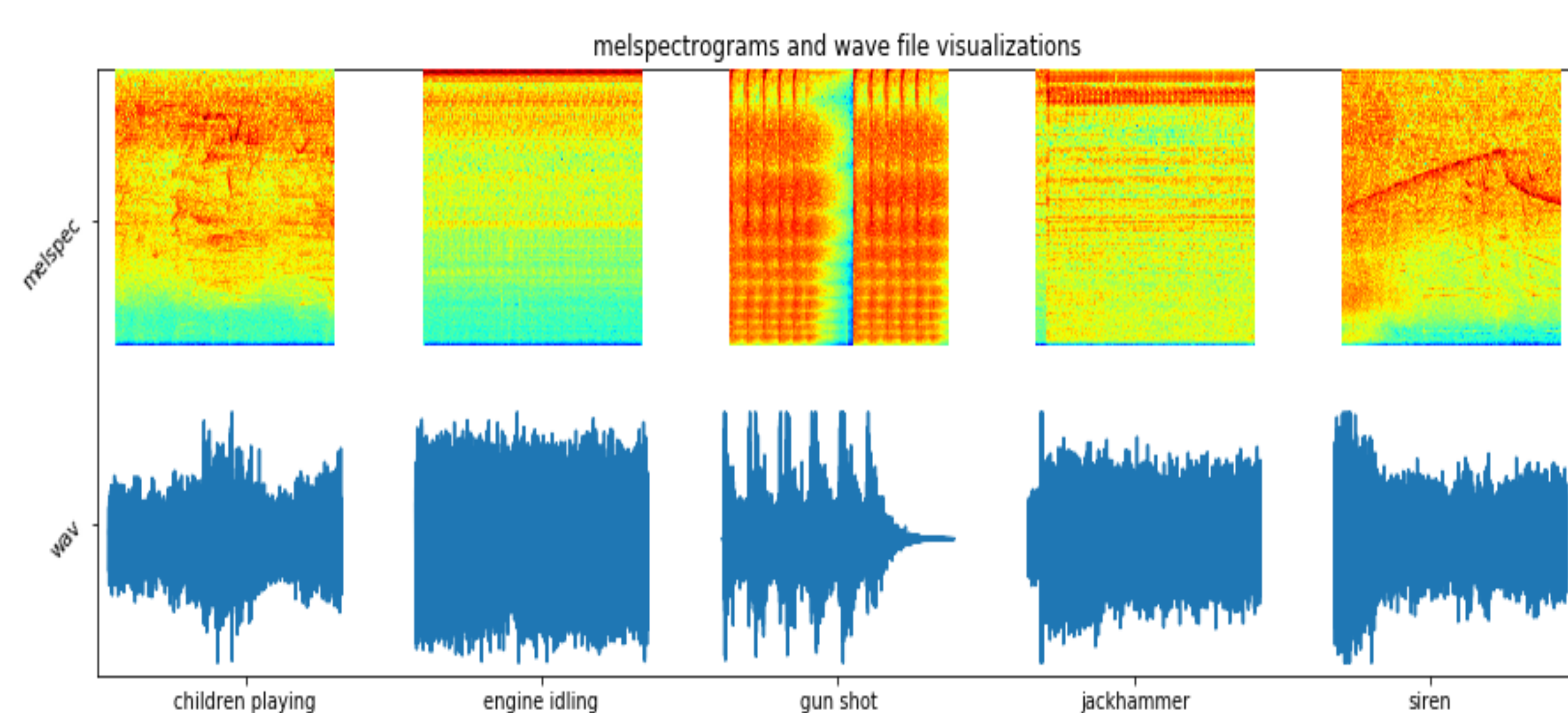
Histogram of Data Distribution Among the 10 Classes

- The UrbanSound8K dataset consists of 8,732 labeled sound excerpts that are $\leq 4s$ in duration.
- 10 Classes include: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music

Sample images



The images in the top row show log-scaled Mel-spectrograms extracted from audio files originally in .wav. Images in the bottom row show the plots of the original .wav files.



Methodology

- Used pre-partitioned UrbanSound8K dataset, using cross validation folds 1-7 for training and 8-10 for testing
- Apply following feature extraction to the training and test set using Librosa audio package:
 - Mel-Frequency Cepstral Coefficients (MFCC):** Coefficients derived from a cepstral representation of the audio clip
 - Chromagram:** Pitch class profiles. They capture harmonic and melodic characteristics within the music
 - Mel-scaled spectrogram:** Psychoacoustic scales that capture the distances from low to high scale frequency
 - Contrast:** Difference between parts of a sound or different instrument sounds
 - Spectral Contrast:** Represents the strength of spectral peaks and valleys in each a sub-band as contrast distribution
 - Tonnetz:** Representation of tonal space
- Compare performances among different neural network architectures and machine learning models

Comparison of Models

We'd like to compare several different neural network architectures and machine learning models to then choose the one with the best performance to do our classification.

- K-Nearest Neighbor:** A supervised learning algorithm used for classification and/or regression.
- Deep Neural Network:** A deep convolutional neural network that utilizes 3 hidden layers. This is a feedforward network, in which data flows from the input layer to the output layer without looping back.
- Convolutional Neural Network:** A fully-connected convolutional neural network that utilizes 1 hidden layer. It is a regularized version of a multilayer perceptron.

Results

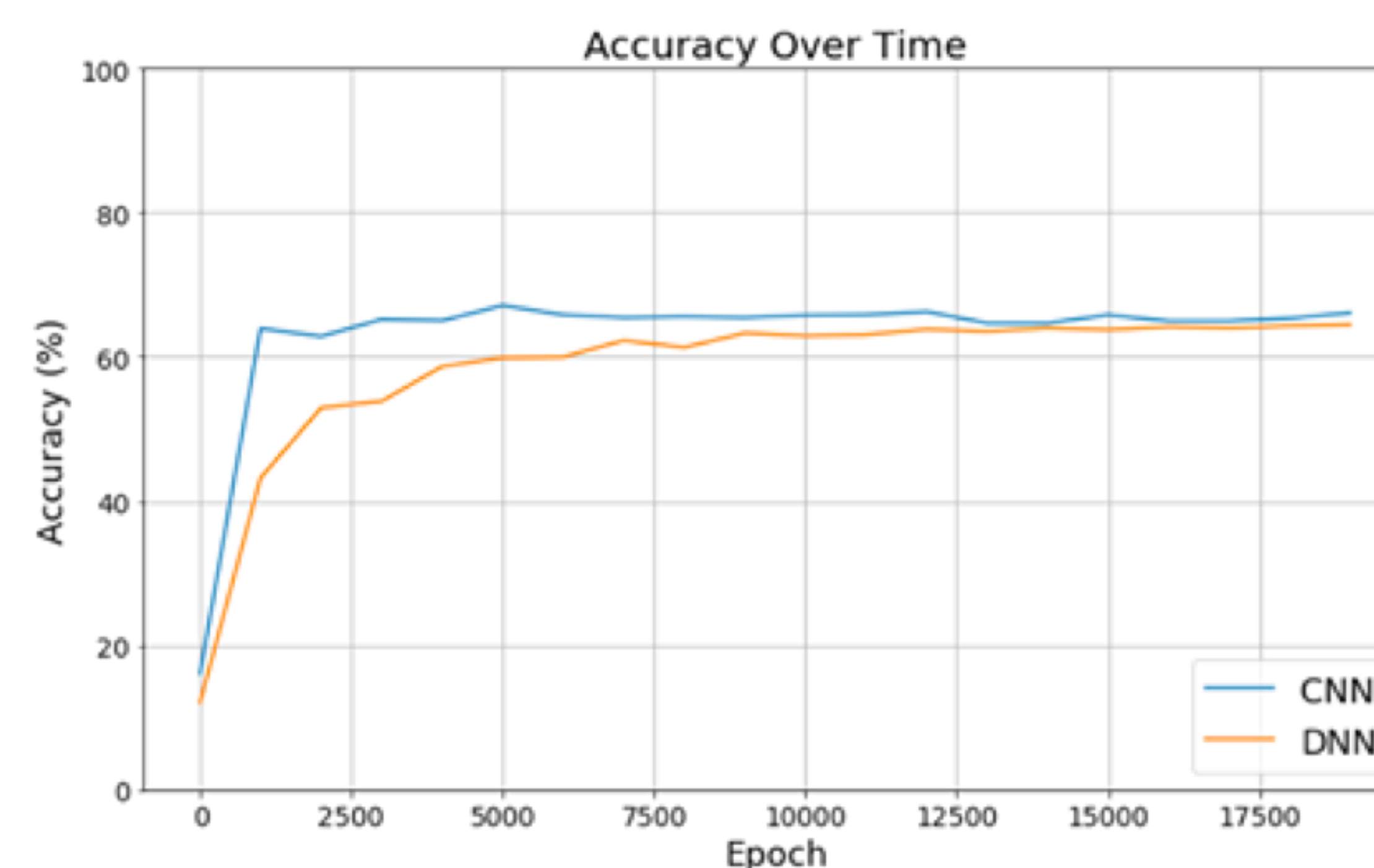


Fig. 1 - The accuracy of our networks over time

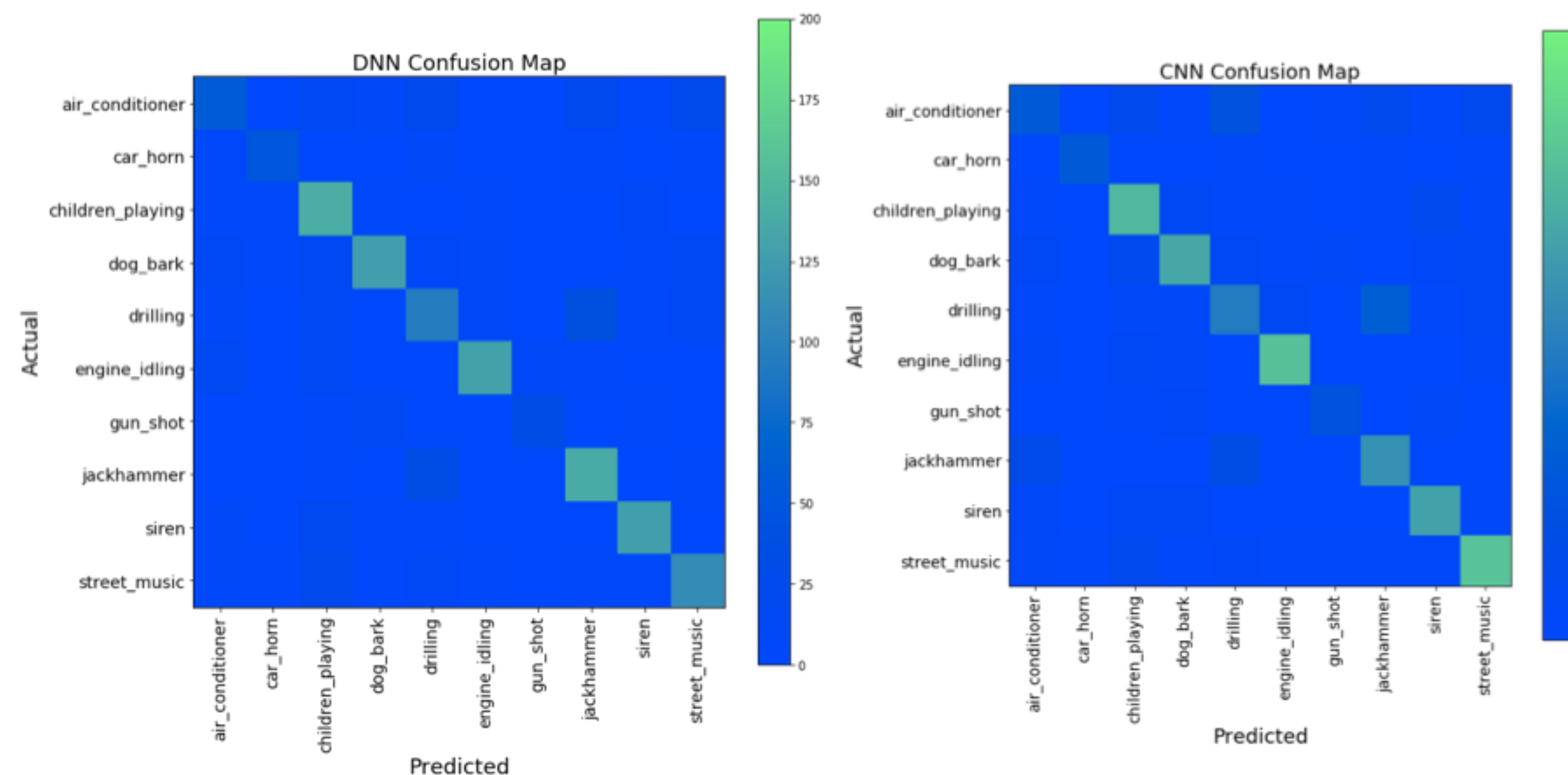


Fig. 2 – Confusion matrices for the DNN(Left) and CNN(Right)

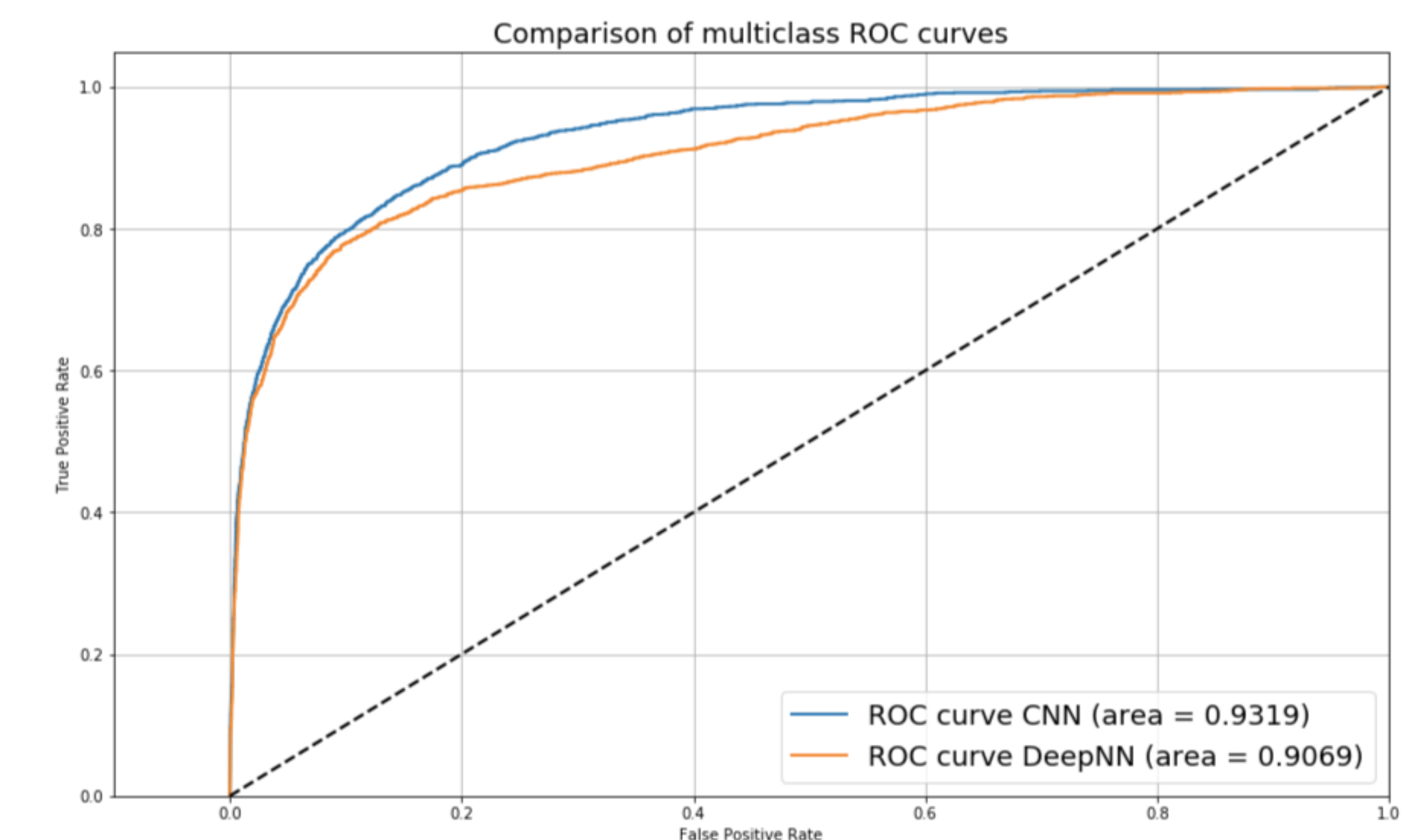


Fig. 3 – ROC curves for our models

Model	ACC	AUC
KNN	55.6%	84.4%
DNN	64.5%	90.6%
CNN	65%	93.1%

Tab. 1 – Final comparison results of model performance. Evaluation metrics used were accuracy and area-under-curve.

Discussion

- We had the best results with the Convolutional Neural Network and the Deep Neural Network
- Overfitting was a big challenge in the model performance, and regularization did not give improvement
- Additional challenge was optimizing speed for extracting features into appropriate inputs for models and networks

Future Work

- Try different methods of feature extraction
- Tailor the neural network architecture to match the extracted features
- Apply class conditional data augmentation to see if it improves importance
- Experiment with established model for performance improvement

[1] Doran, Benjamin. "BenjaminDoran/Urban-Sound-Classification." *GitHub*, 11 June 2017, github.com/BenjaminDoran/Urban-Sound-Classification.

[2] Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. "Music type classification by spectral contrast feature." In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1, pp. 113-116. IEEE, 2002.

[3] J. P. B. Justin Salamon, Christopher Jacoby, "Urbansound8k dataset." [Online]. Available: <https://urbansounddataset.weebly.com/urbansound8k.html#10foldCV>