SPEECH CLASSIFICATION USING CNN, SVM, REGRESSION TREES

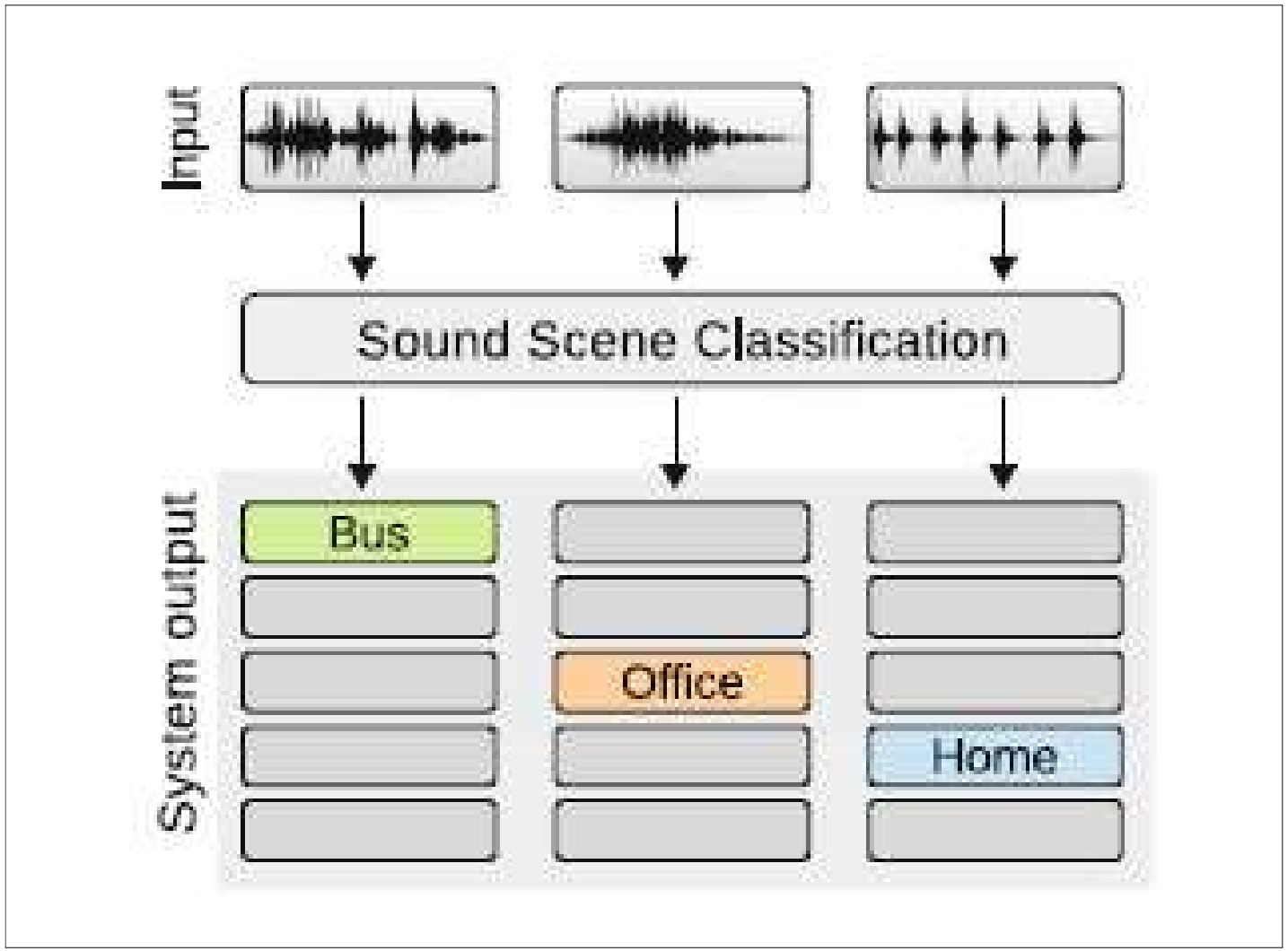
By
Clyde Clarke

Outline

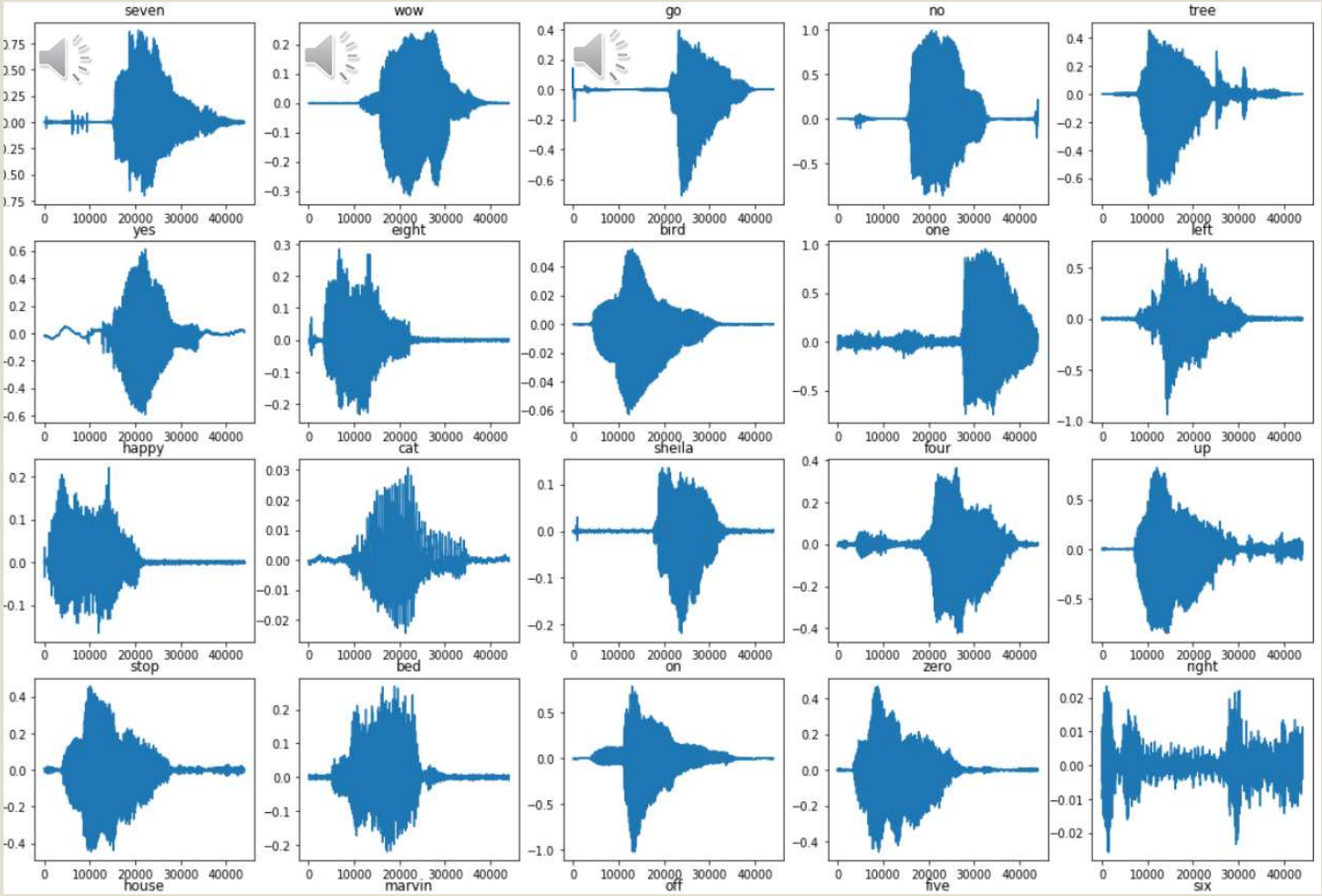
- Problem Statement
  - Dataset
- Sound Classification
- Feature Extraction
  - MFCC
  - STFT
  - WAVELET
- Classification Methods
  - SVM
  - Regression Trees
  - CNN
- Evaluation

The Speech Dataset

- Speech Commands Dataset
- Contains 15000 labeled sound excerpts (1s) of real field-recording speech from 30 classes : (1).nine, (2).eight, (3).one, (4). zero, (5). bird, (6). shiela, (7). night, (8) siz, (9) stop, up (10.)… etc.
- Sampling Frequency 44 KHz
- Sample Duration ~ 1s
- Sound Samples 15000



Sound Classification

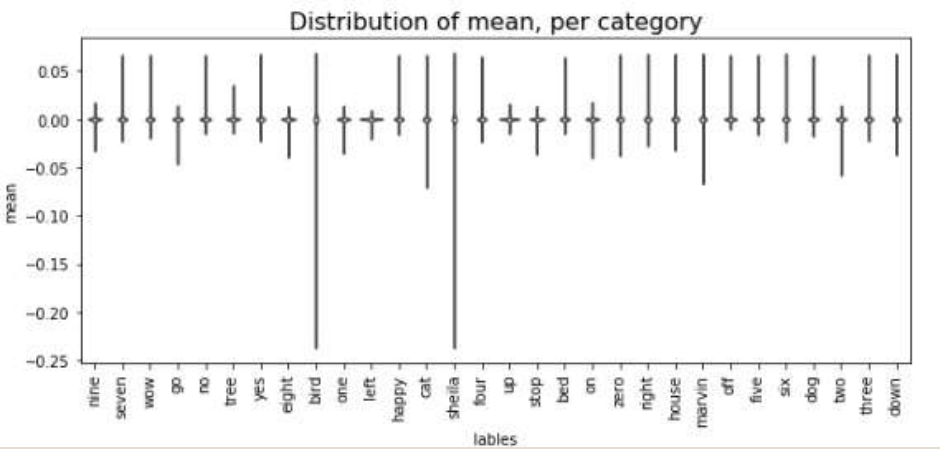


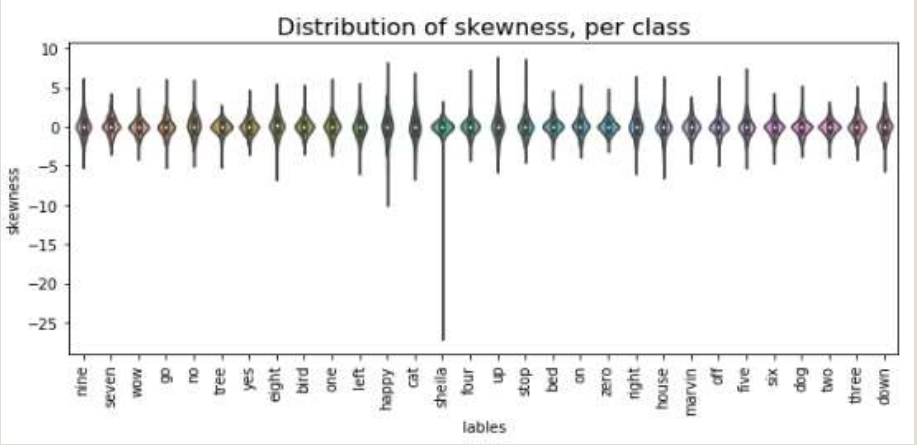
SOUND CLASSES

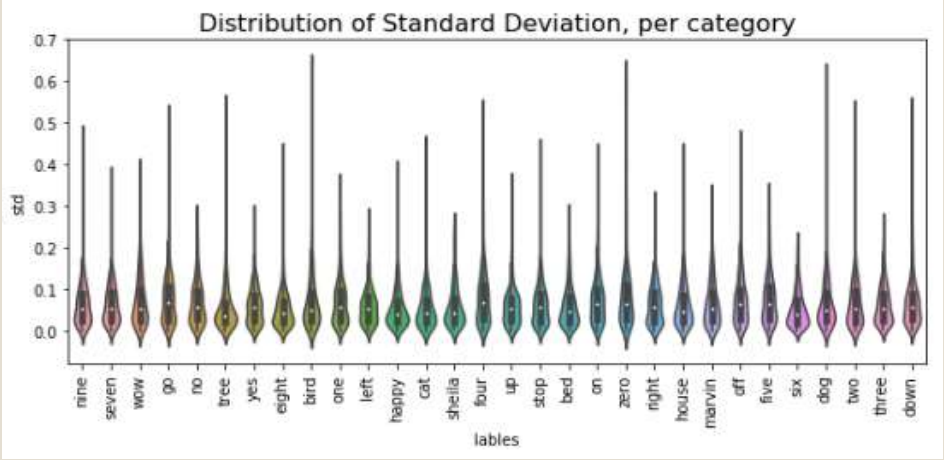
Sound samples of 1 second in duration

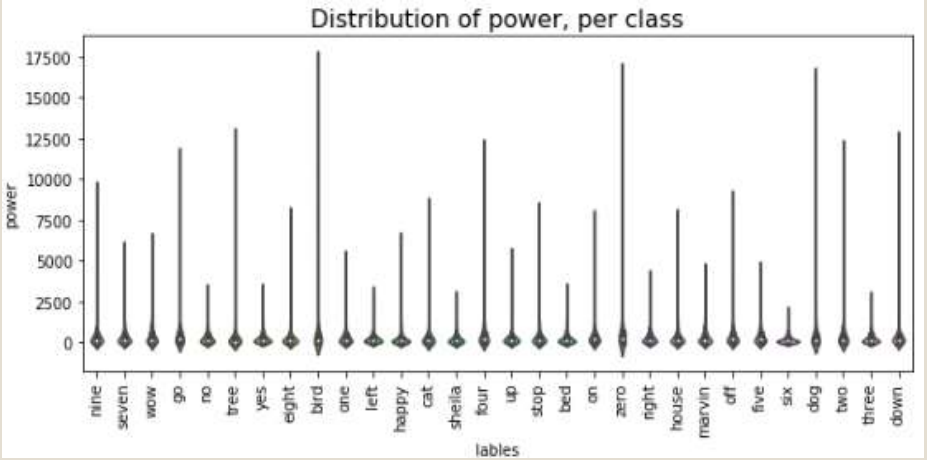
Each audio class is 44100 sample

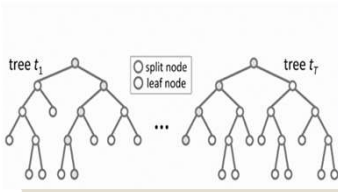
Supervised Learning Features

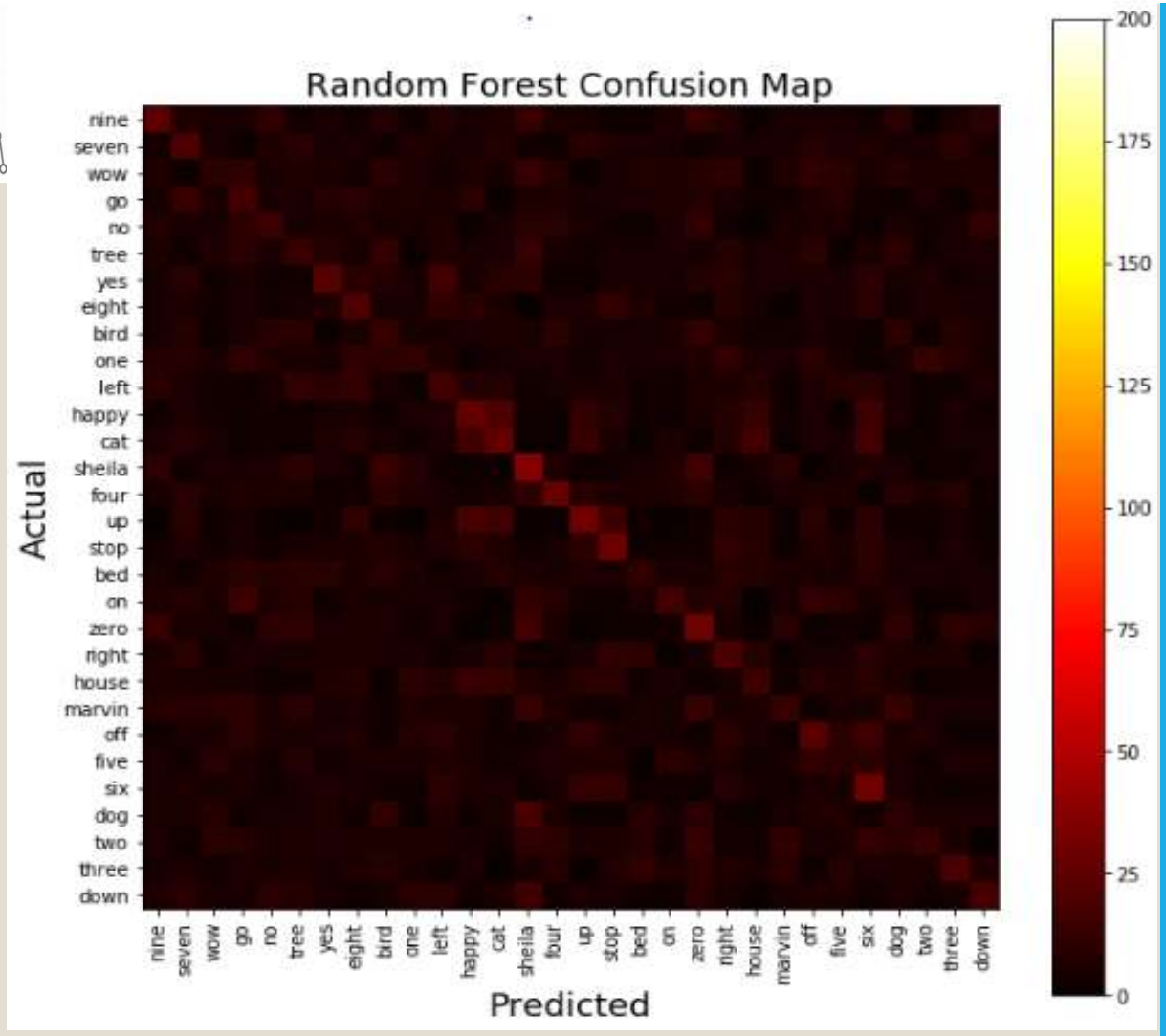






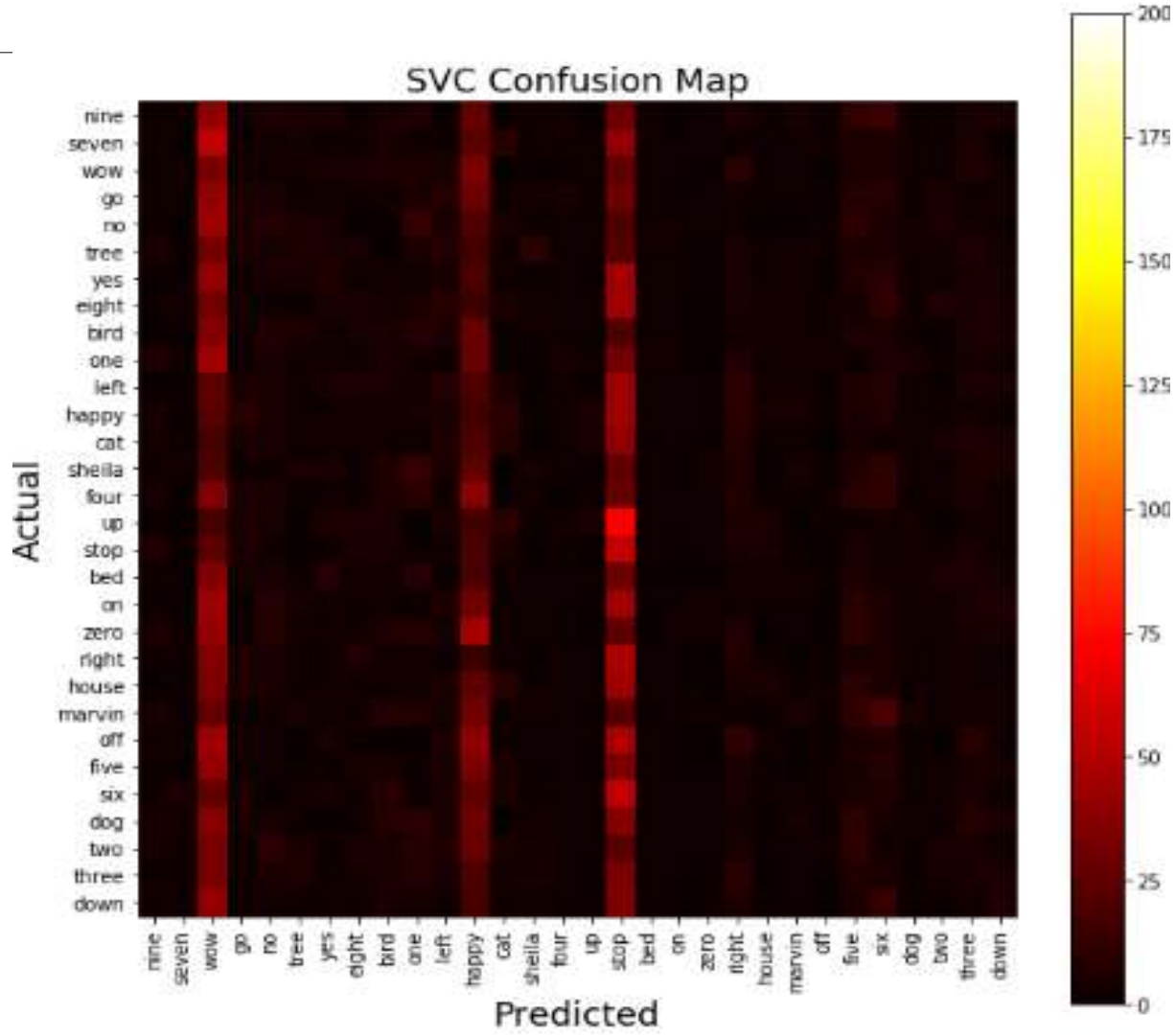






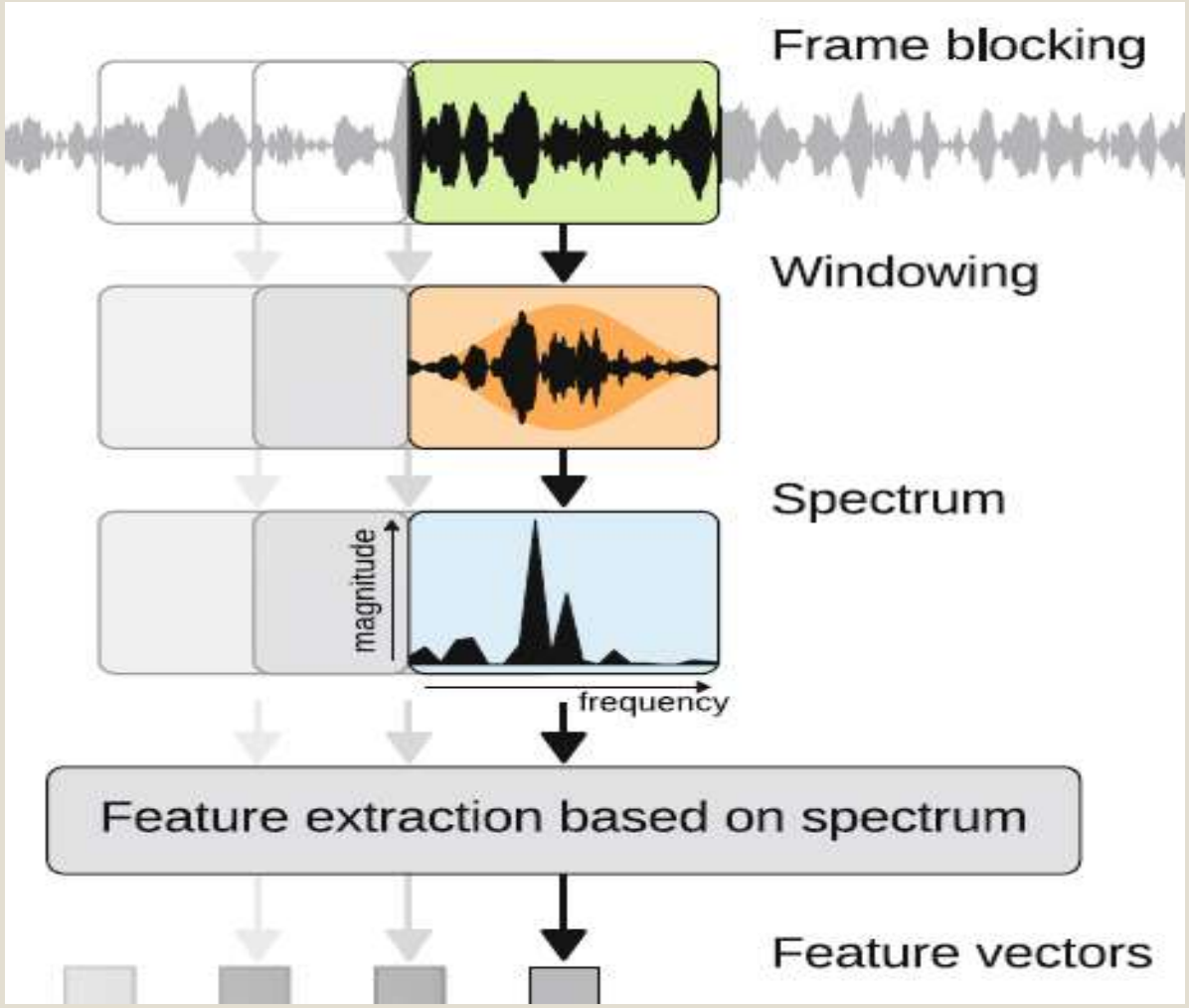
Random Forest

- Random forests (RF) are a combination of tree predictors
- Each tree depends on the values of a random vector sampled in dependently
- The generalization error depends on the strength of the individual trees and the correlation between them
  - Accuracy
    - .11



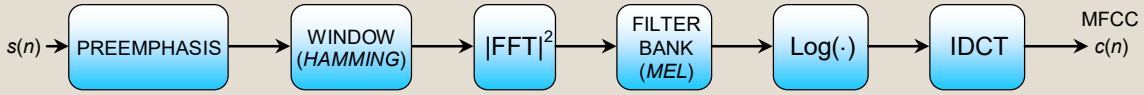
SVM

- Large Margin Classifier
- Maximizes the distance between decision boundary and support vectors
- Kernel
  - Radial Basis Function
- Gamma
  - .08
- Accuracy
  - .11

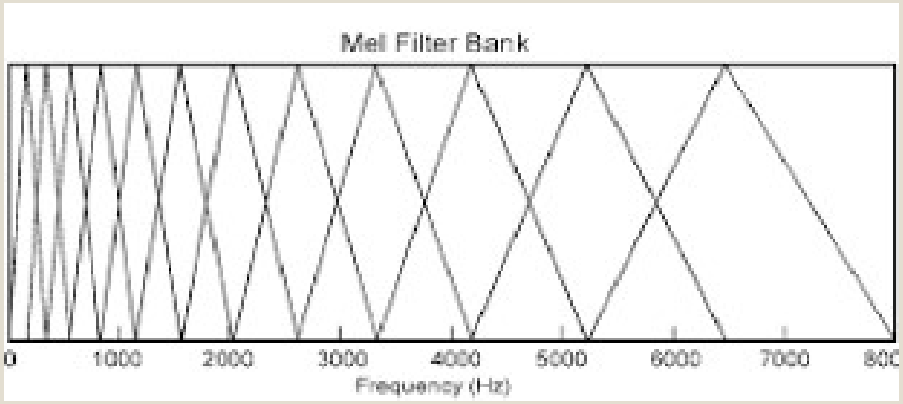


Feature Extraction

- Frame Blocking
  - Segment signal into regions that retain small scale changes
- Windowing
  - Done to prevent gibbs phenomena at frame edges that would corrupt spectral analysis
- Spectrum
  - Any number of techniques can be used to extract features from signal
    - DWT
    - FFT
    - DCT



- Preemphasis: compensates for spectral tilt (speech production/microphone channel)
- Windowing: suppression of transient effects in short-term segments of signal
- $|FFT|^2$: energy spectrum (phase is discarded)
- MEL Filter bank: MEL scale – models logarithmic perception of frequency in humans; triangular filters – dimensionality reduction

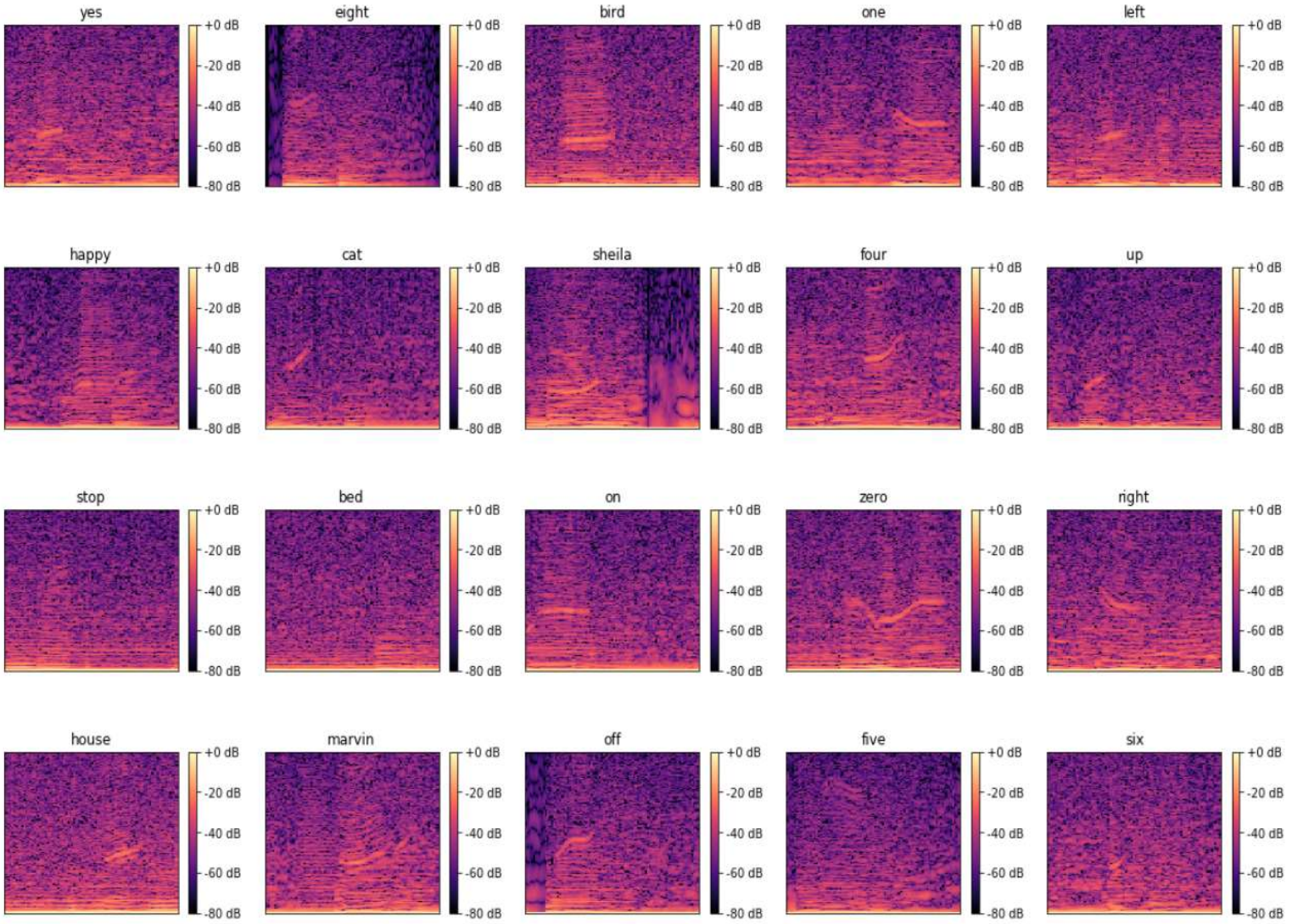


MFCC Features

Pre emphasis filter performs lowpass at Nyquist frequency of signal

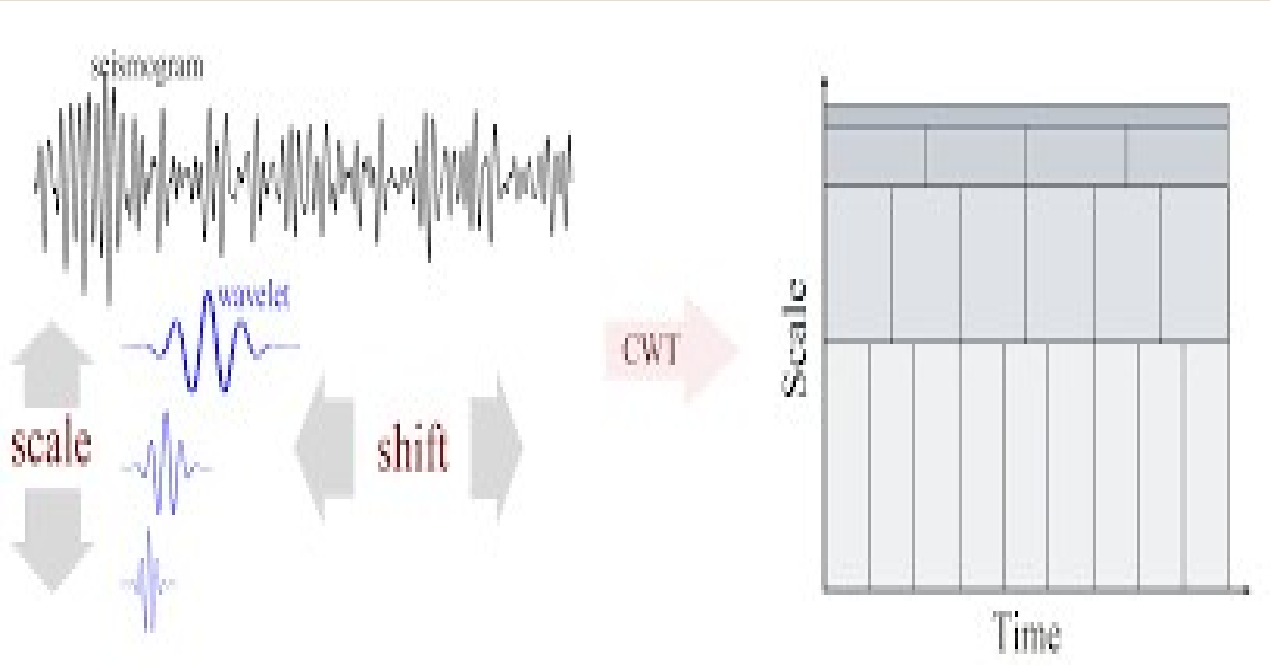
Short Time Fourier transform is performed on the windowed signal

Logrithmic filter bank is used that models the human perception of sound



MFCC FEATURES

Figure shows a representative sample from 16 distinct sound classes



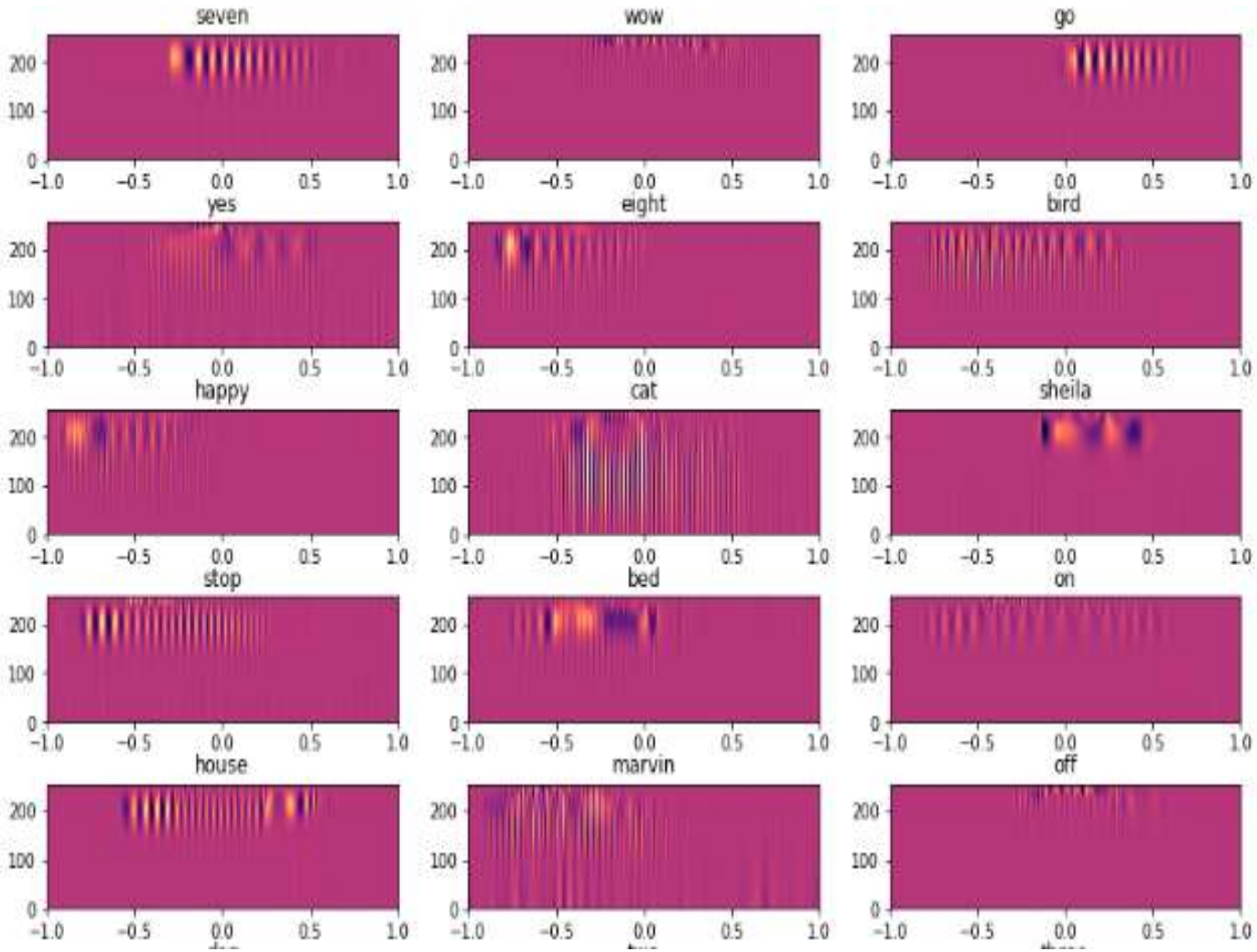
$$(W_a f) = \int f(x)\psi_{a,b}(x)dx$$

$$with \quad \psi_{a,b}(x) = {}^{1}/_{\sqrt{a}}\, \psi\left(\frac{x-b}{a}\right)$$

Wavelet Transform

Continuous Wavelet Transform

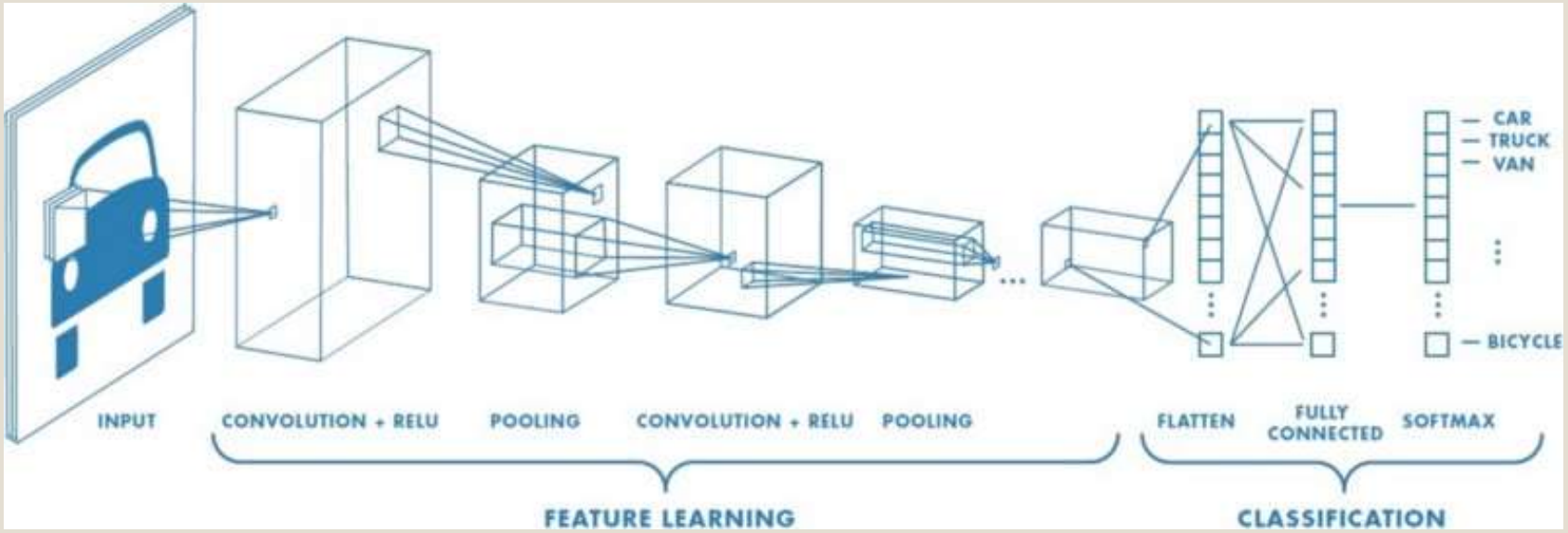
- Decomposes signals with (wavelets) orthogonal basis functions
- Scalogram is computed by shifting and scaling (ie stretching) wavelet function
- Similar to Fourier Transform



WAVELETS

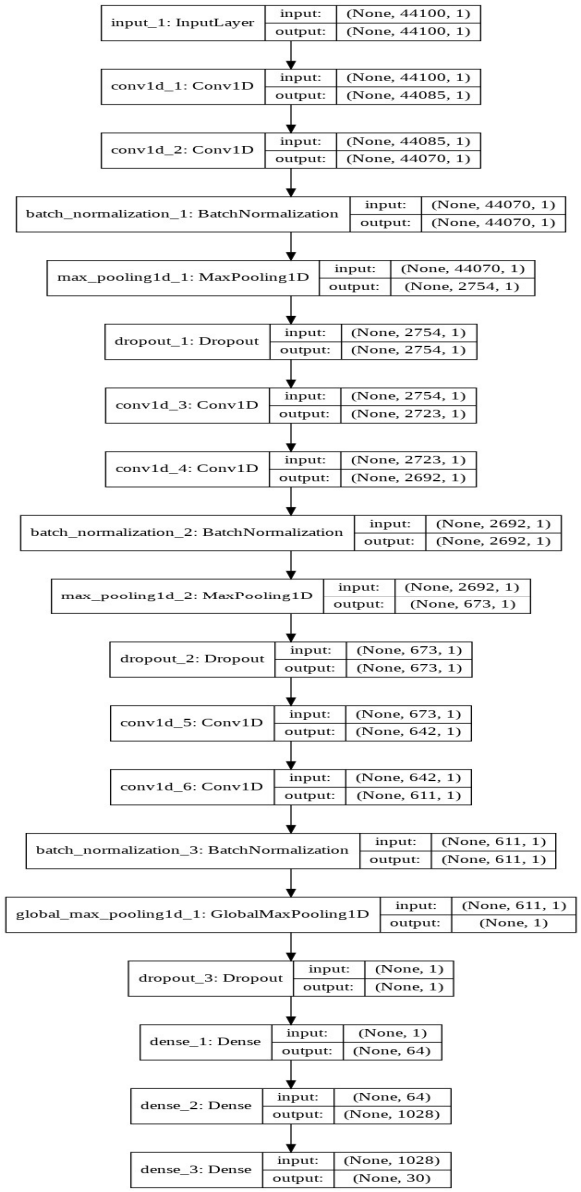
Continuous Wavelet Transform (CWT) using Mexican Hat Wavelet

Convolutional Neural Network Architecture

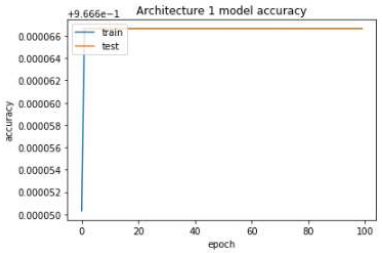


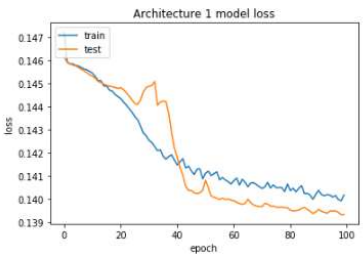
Training and Processing

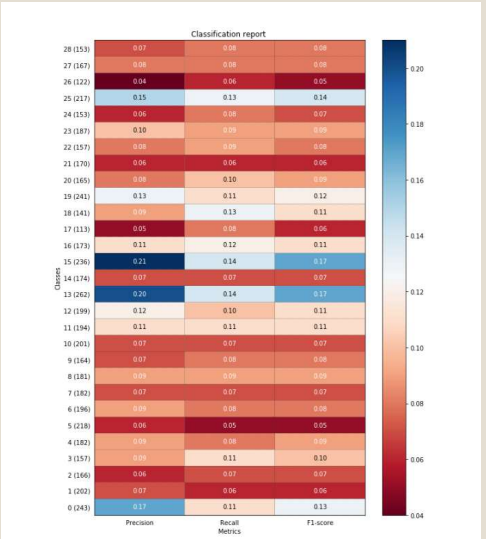
- Due to variable sample length wave files were zero padded
- Training was done using an 60/40 split of the data
- Network was trained using adam optimization
- A batch size of 256 and 100 epoch was used to train five convnet architectures



CNN ARCHITECTURE 1

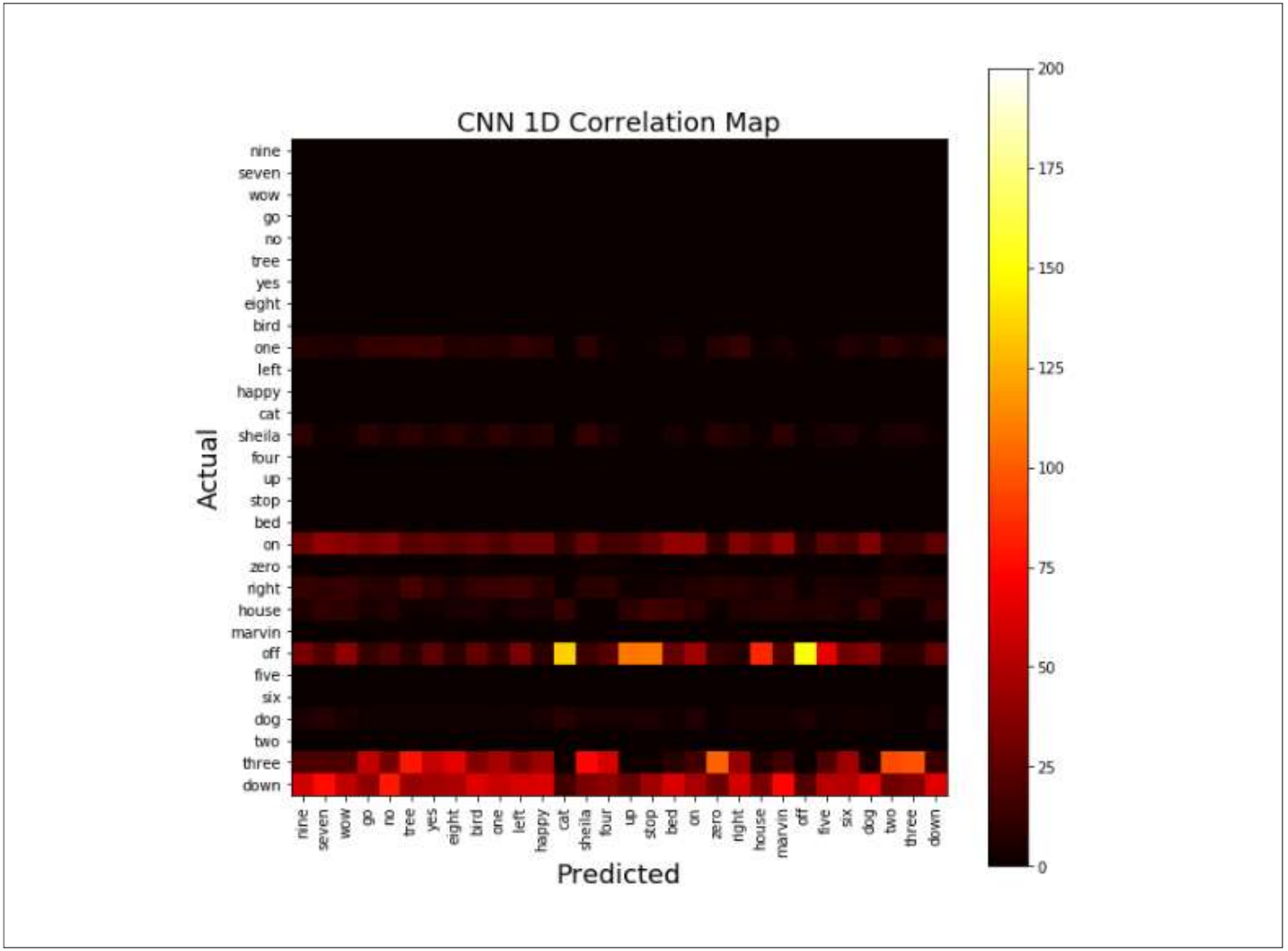






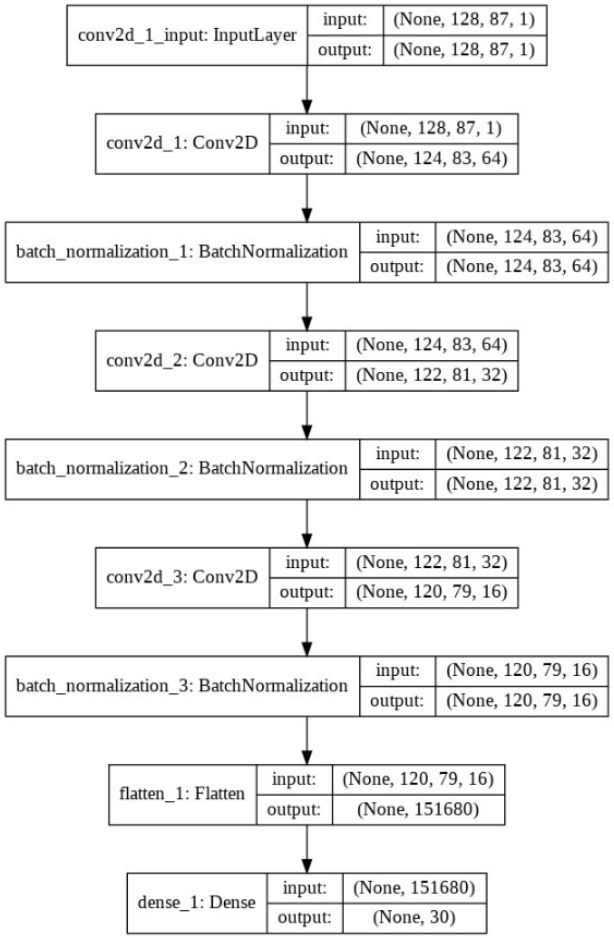
Architecture 1

- Training of the model for both validation and test data shows poor performance
- The classification report also shows poor precision, recall and f-scores across all classes



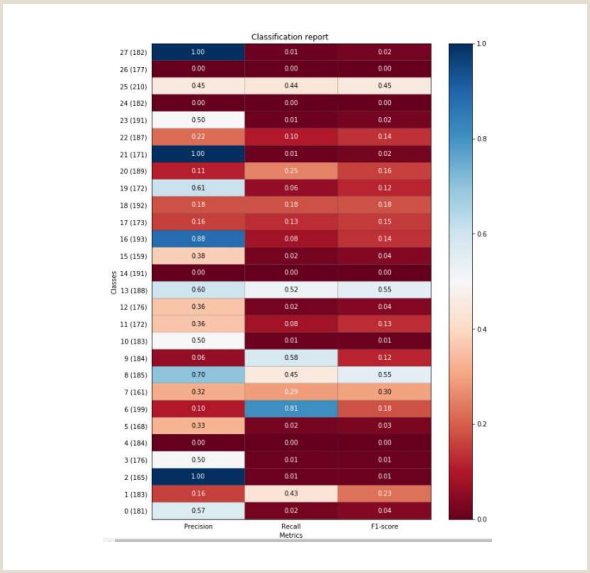
Correlation Map

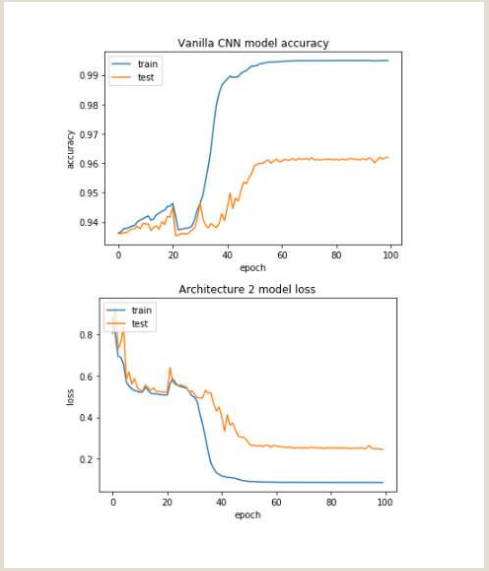
- The correlation map of the 1D model shows poor performance as it mistakes classes
- The model accuracy is merely 10%



ARCHITECTURE 2

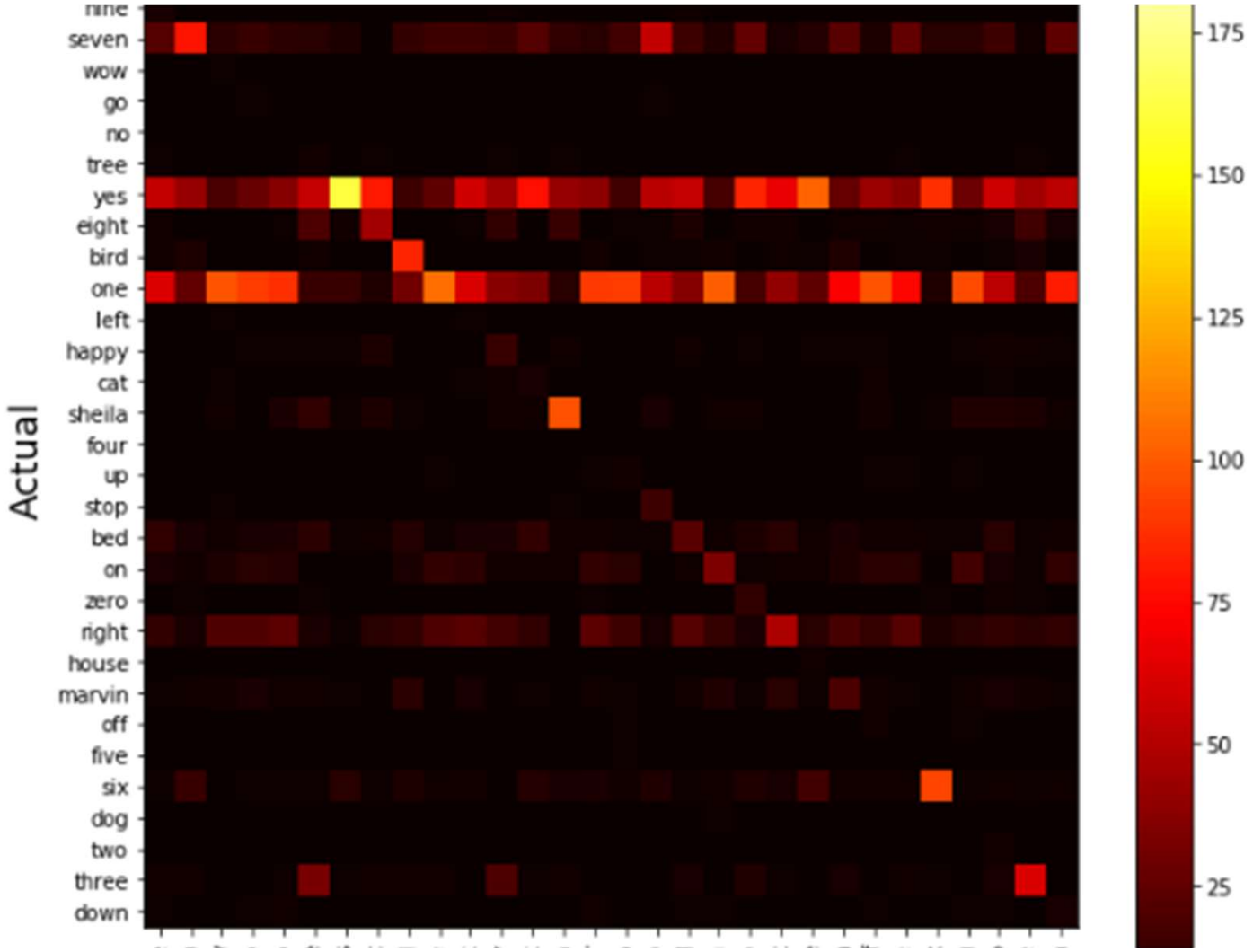
Input features for this architecture are 2D MFCC with 200 filters





Performance

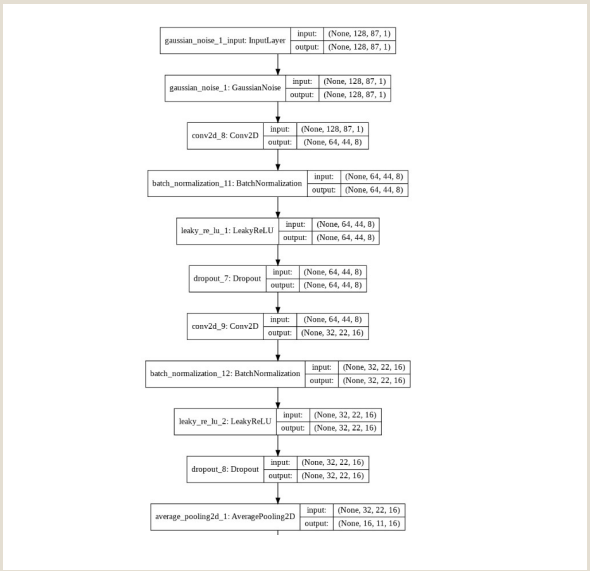
- Trained model has increasing loss and increasing accuracy over 100 epochs
- The performance however is still bad for this model
- Accuracy
  - 12%

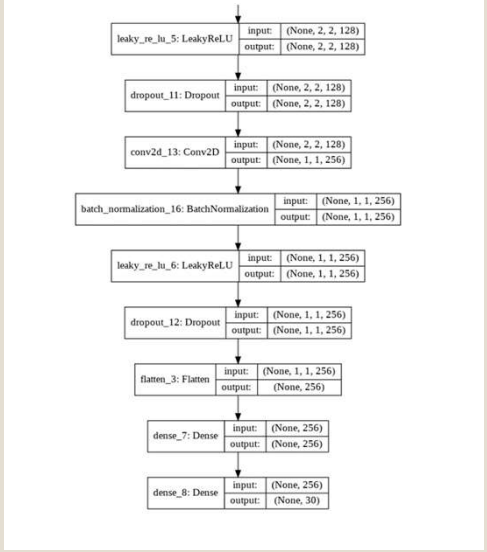


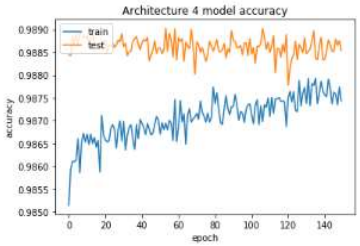
Correlation Map

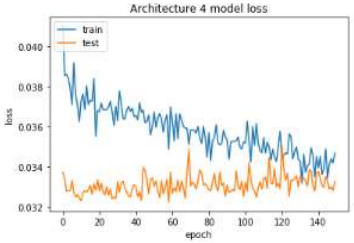
- Training of this model shows that classification of yes and one had high correlation with all other audio classes in the dataset
- This is probably due to not having enough features ie convolutional layers in the model

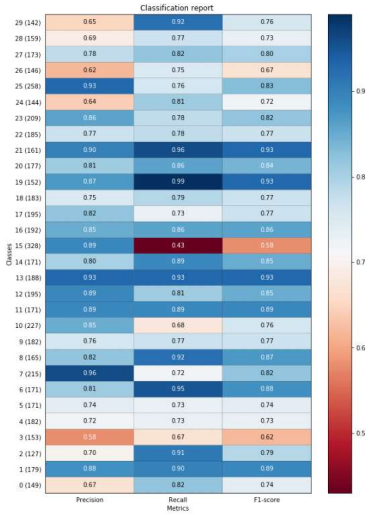
Architecture 3





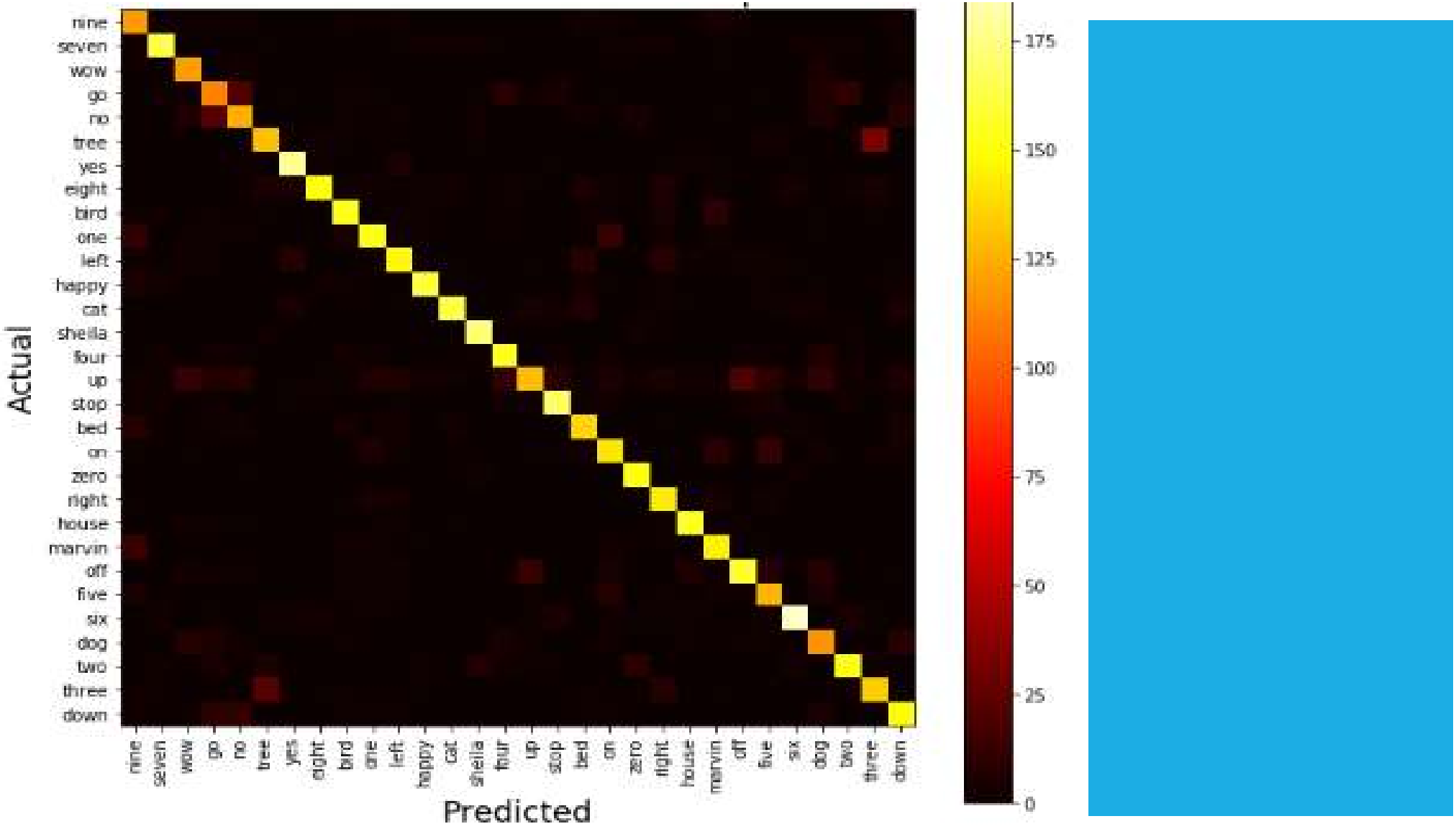






Deep CNN 2D

- Training and Classification Report for Deep CNN with 20+ layers
  - Layers include
    - 2D Convolution
    - Batch Normalization
    - Dropout
    - Dense Layers
- Training Accuracy
  - 96%
- Model score 81%


Actual
nine
seven
wow
go
no
tree
yes
eight
bird
one
left
happy
cat
sheila
four
up
stop
bed
on
zero
right
house
marvin
off
five
six
dog
two
three
down
Predicted
175
150
125
100
75
50
25
0

Conclusions

- Supervised Learning techniques performed poorly. This is probably due to the feature space chosen in addition to the large variability in the dataset
- 1D Convolutional Neural Networks and newtorks with fewer layers also showed poor performance
- Deep 2D Convolutional Neural Network with MFCC features performed the best giving 81% accuracy
- The CWT coefficient features for the Deep 2D CNN also provided low accuracy
  - This is probably due to having to restrict the signal by using interpolation to reduce time sample size
- Hyperparameter tuning
  - Hyperperameter tuning was used via the hyperopt library in python
  - After optimization of the hyperparameter space the models were run.
  - Issues with the choice of which parameters to vary as well as memory limitations restricted an exhaustive search of the parameter space