Deep Learning

MUSIC GENRE CLASSIFICATION

Dinmukhammed Sagatbekov

Presentation outline

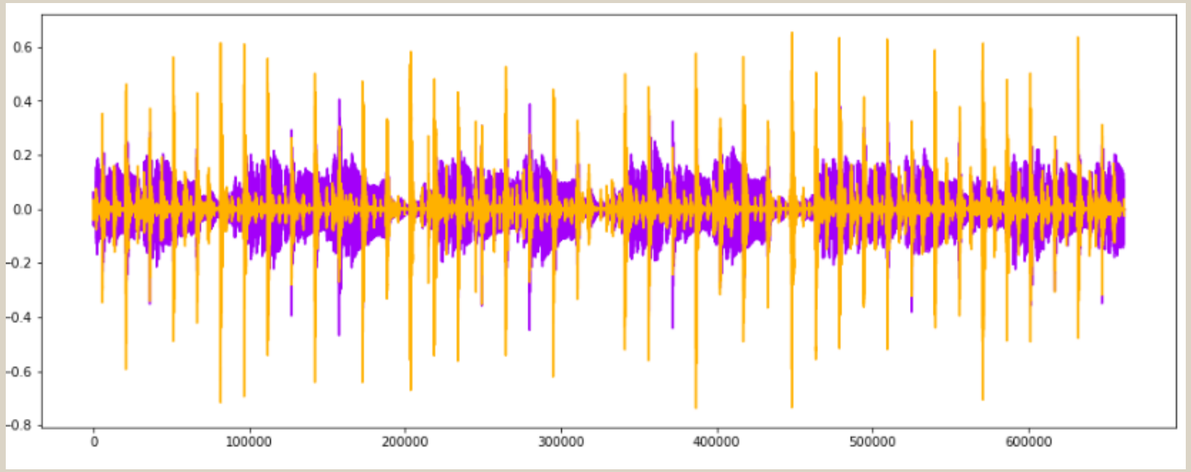
- Dataset description
- Feature extraction
- Model description -> Results
- Conclusion

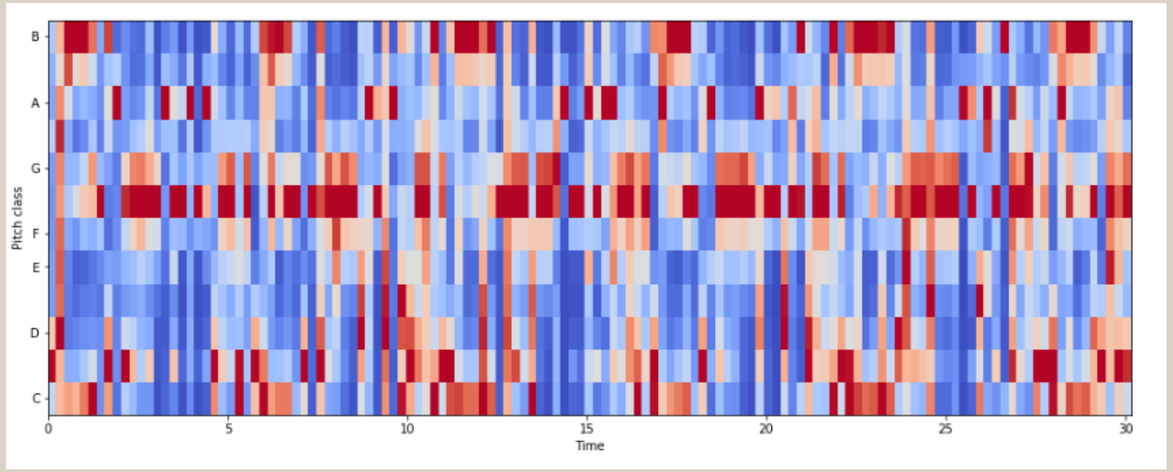
Dataset description

- The famous GTZAN dataset
- Benchmark for music classification
- 1000 songs for 10 genres
- Heavy artist repetition, which often ignored during dataset split
- The labels not 100% correct

Features

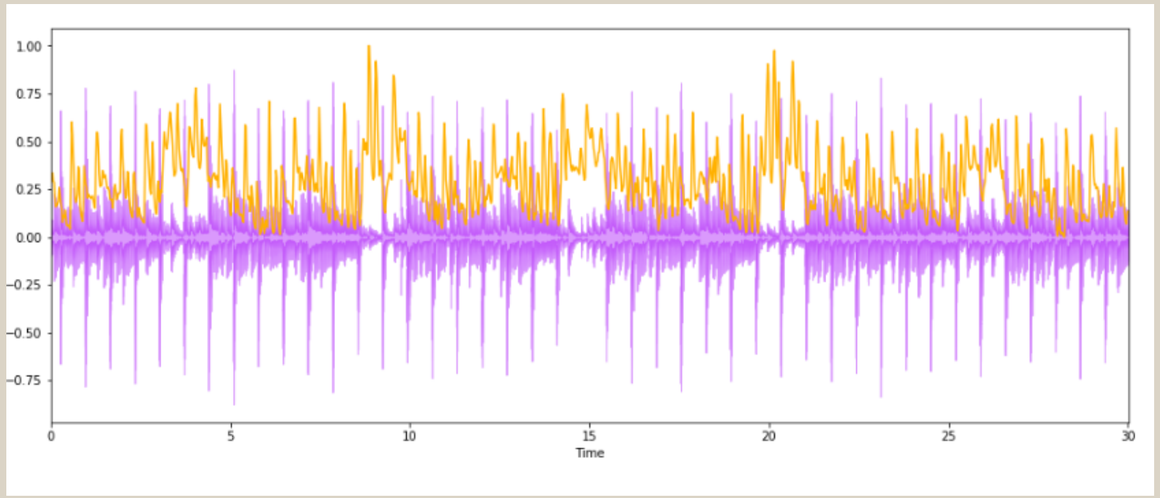
- Zero Crossing Rate - information about the rate at which the signal changes its sign, or crosses zero, over a given period
- Harmonics (graph 1)
- perceptual features are characteristics of the sound that are related to human perception. These features can include loudness, pitch, duration, and timbre (graph 1)
- Chroma features - derived from the chromagram, which is a representation of the energy distribution of musical pitch classes (notes) in an audio signal (graph 2)

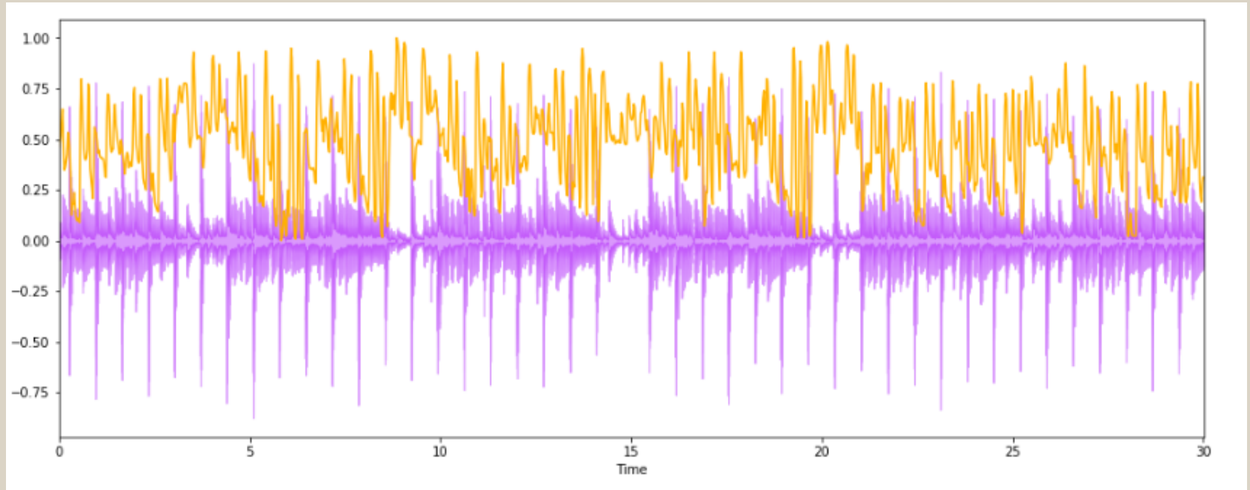




Features

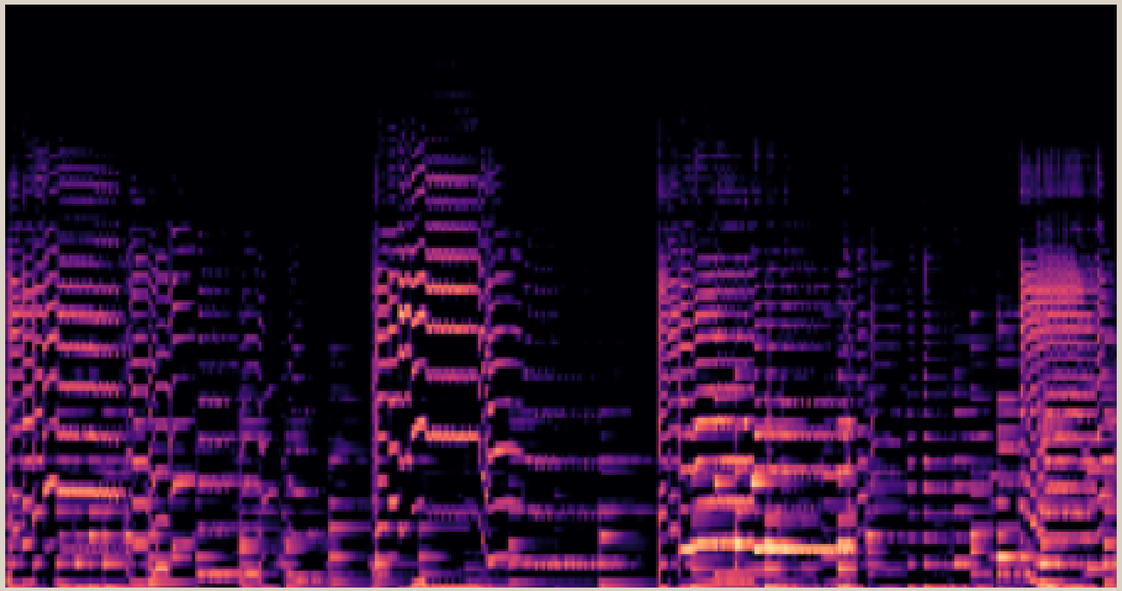
- Tempo BMP (beats per minute)
- Spectral Centroid - indicates where the "centre of mass" for a sound is located. Centroids (yellow) along waveform (graph 1)
- Spectral Rolloff - a measure that provides information about the shape or steepness of the spectral content of a signal. Spectral rolloff is often employed in the field of audio and music processing to characterize the frequency distribution of a sound (graph 2)

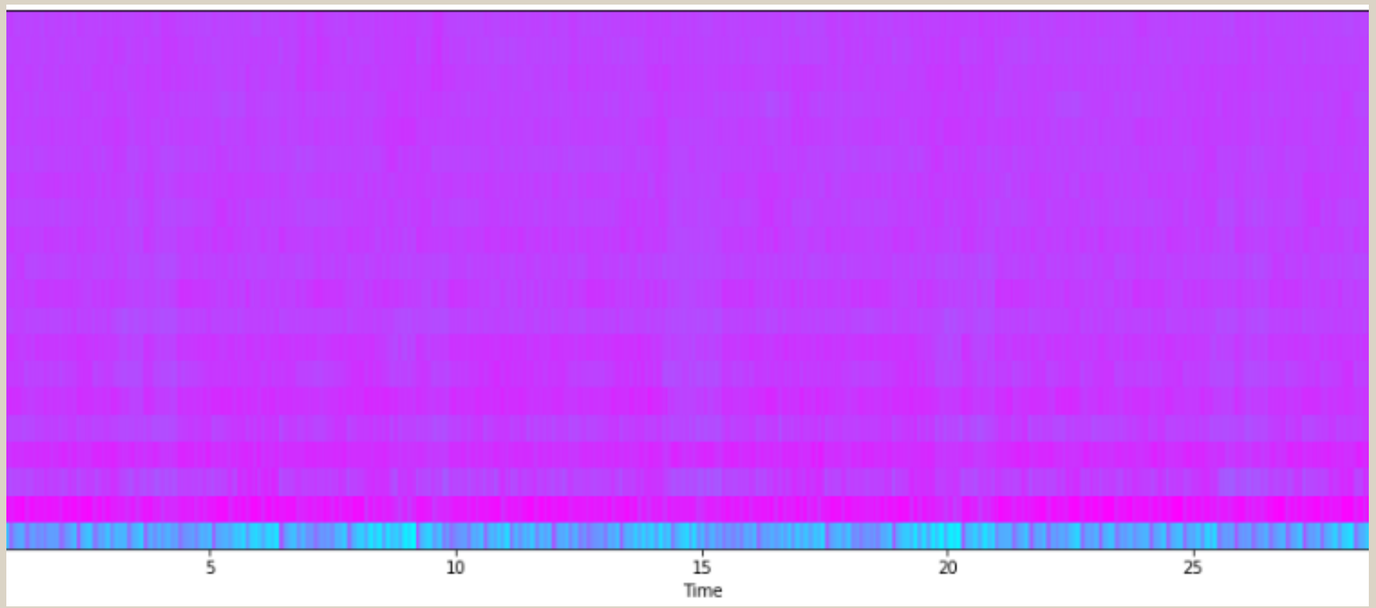




MFCC Feature

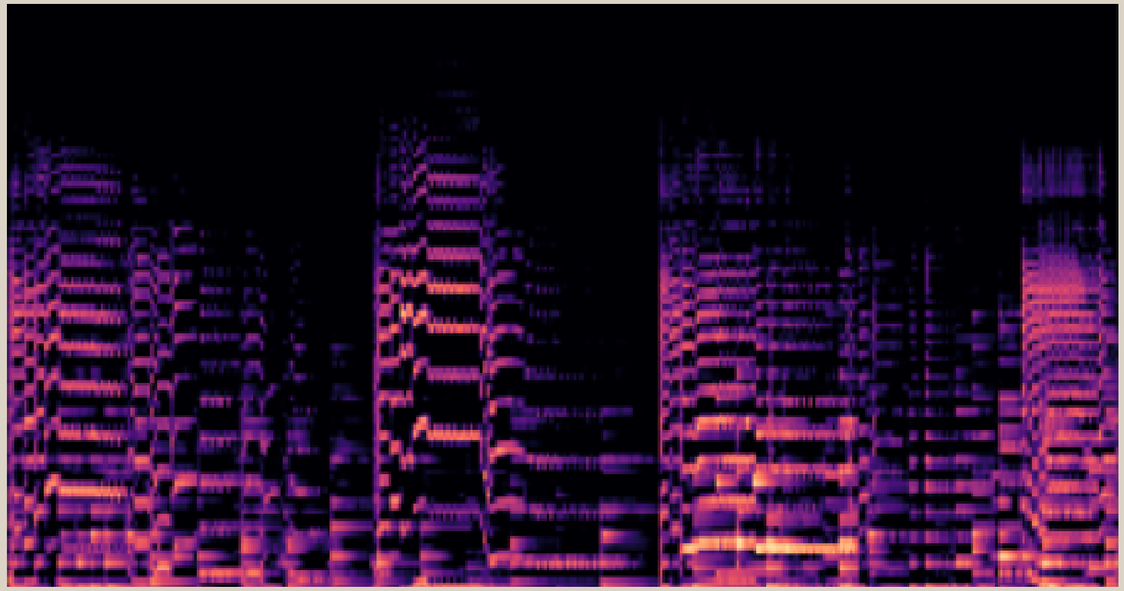
- Mel-Frequency Cepstral Coefficients (MFCCs) - coefficients that collectively represent the short-term power spectrum of a sound signal (graph 2)
- Widely used in audio signal processing, speech processing, and music analysis
- A Mel-frequency spectrogram (MFS) is a representation of the short-term power spectrum of an audio signal that takes into account the nonlinear human perception of sound frequencies (graph 1)

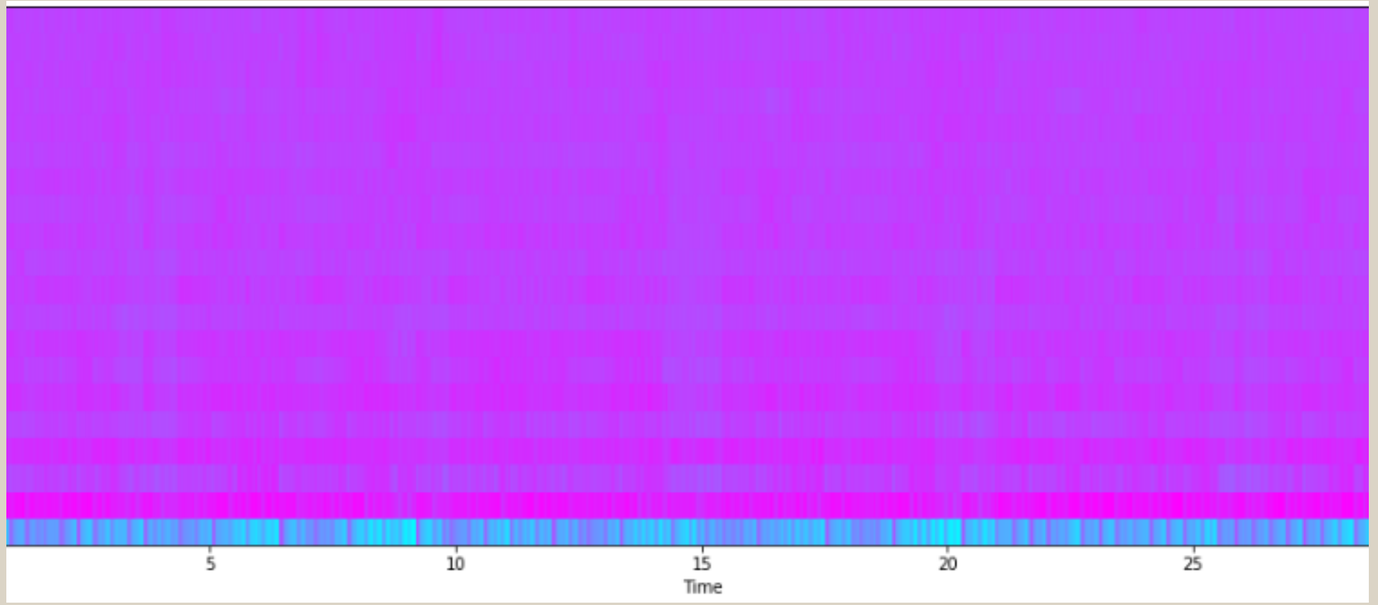




MFCC Feature

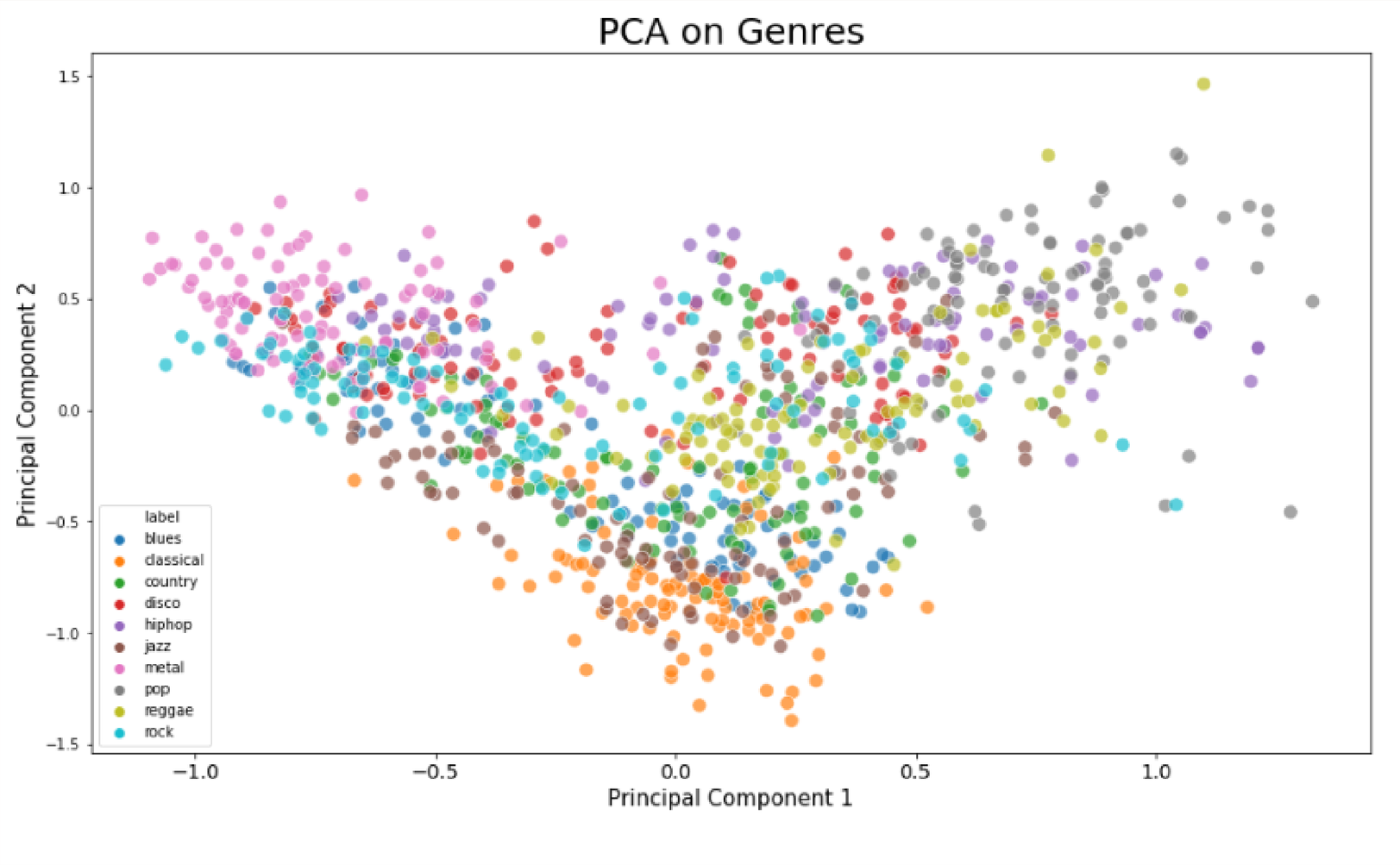
- The audio signal is divided into short frames, often on the order of 20-40 milliseconds each
- To ensure continuity in the analysis and prevent loss of information at frame boundaries, frames are typically overlapped
- MFCC Mean - The MFCC mean is the average value of the MFCC coefficients over a certain duration or set of frames -> 20 MFCC means (for each 1.5 sec)
- MFCC Variance- represents the spread or dispersion of the MFCC values. It measures how much individual MFCC values deviate from the mean -> 20 MFCC vars





Data visualization with PCA

- Dimensionality reduced to transform high-dimensional data into a new coordinate



Data Separation

- 2 Datasets - 30s and 3s
- 30s data split 70% to 30%. Stratified Sampling
- How to split 3s dataset?
  1) Whole song (3s x 10) in one of the sets -> overfitting problem, low test accuracy
  2) Stratified Sampling over whole dataset -> same song in train and test sets

- Many projects use second method for 3s dataset -> High accuracy. Possibility of memorization of song

Tested simple models

30s dataset

```
Accuracy Naive Bayes : 0.53

Accuracy Stochastic Gradient Descent : 0.66

Accuracy KNN : 0.665

Accuracy Decission trees : 0.5

Accuracy Random Forest : 0.68

Accuracy Support Vector Machine : 0.705

Accuracy Logistic Regression : 0.655

Accuracy Neural Nets : 0.715
```

3s dataset 1st version

```
Accuracy: 0.37629

Accuracy: 0.44007

Accuracy: 0.44174

Accuracy: 0.36361

Accuracy: 0.40968

Accuracy: 0.46978

Accuracy: 0.46745

Accuracy: 0.47145
```

3s dataset 2nd version

```
Accuracy: 0.5

Accuracy: 0.63614

Accuracy: 0.81832

Accuracy: 0.63664

Accuracy: 0.8003

Accuracy: 0.75125

Accuracy: 0.69069

Accuracy: 0.62613
```

Fully Connected Neural Network

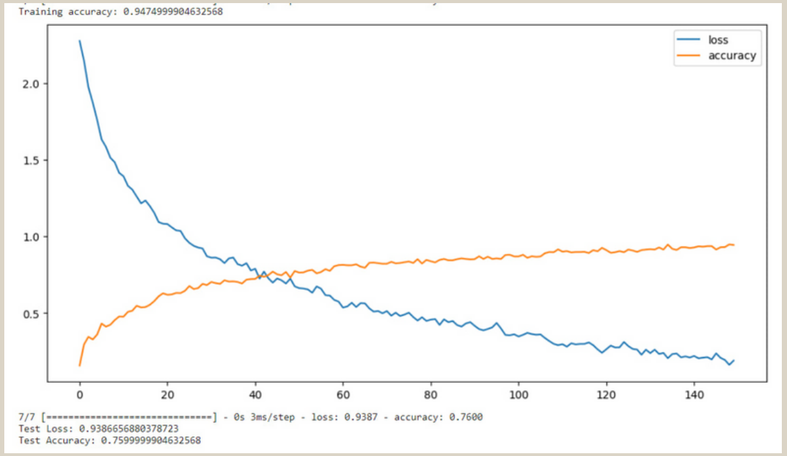
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 1024)              59392

 batch_normalization (Batch  (None, 1024)              4096
 Normalization)

 dropout (Dropout)           (None, 1024)              0

 dense_1 (Dense)             (None, 512)               524800

 batch_normalization_1 (Bat  (None, 512)               2048
 chNormalization)

 dropout_1 (Dropout)         (None, 512)               0

 dense_2 (Dense)             (None, 256)               131328

 batch_normalization_2 (Bat  (None, 256)               1024
 chNormalization)

 dropout_2 (Dropout)         (None, 256)               0

 dense_3 (Dense)             (None, 128)               32896

 batch_normalization_3 (Bat  (None, 128)               512
 chNormalization)

 dropout_3 (Dropout)         (None, 128)               0

 dense_4 (Dense)             (None, 64)                8256

 batch_normalization_4 (Bat  (None, 64)                256
 chNormalization)

 dropout_4 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 10)                650

=================================================================
Total params: 765258 (2.92 MB)
Trainable params: 761290 (2.90 MB)
Non-trainable params: 3968 (15.50 KB)
```

Seven layers:

1. Dense layer with 1024 neurons and ReLU activation.
2. Batch Normalization layer.
3. Dropout layer with a dropout rate of 0.5.
4. Dense layer with 512 neurons and ReLU activation.
5. Batch Normalization layer.
6. Dropout layer with a dropout rate of 0.5.
7. Dense layer with 10 neurons and a softmax activation function.

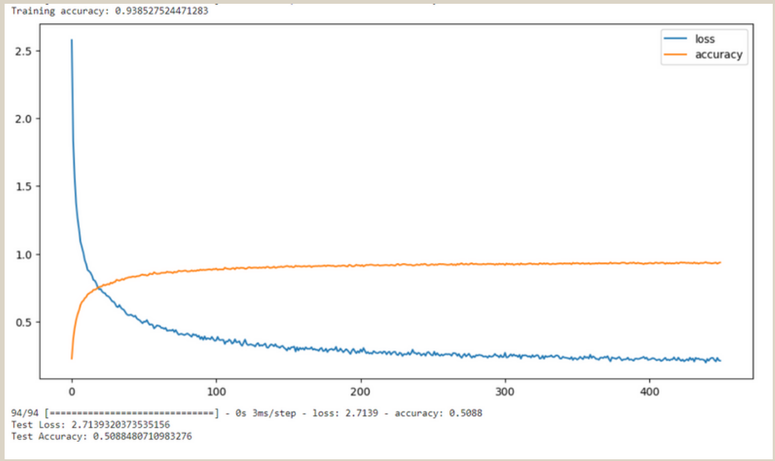
FNN

30s dataset



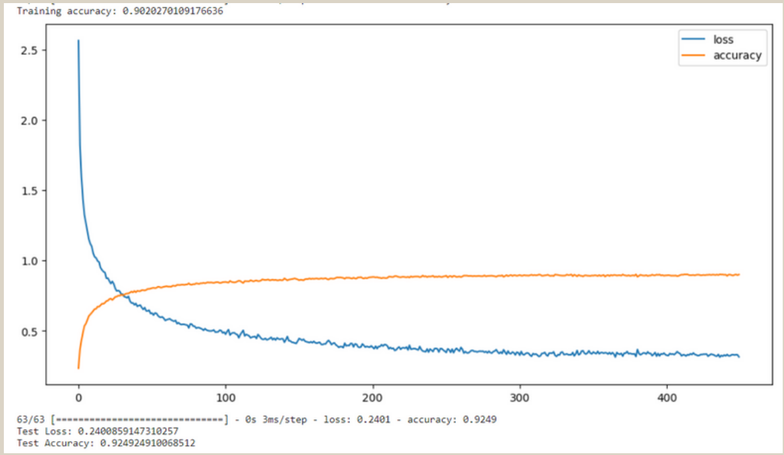
76% accuracy

3s dataset 1st version



50% accuracy

3s dataset 2nd version



92% accuracy

Fully Connected Neural Network (simple)

Layer (type)	Output Shape	Param #
dense_131 (Dense)	(None, 512)	29696
dropout_108 (Dropout)	(None, 512)	0
dense_132 (Dense)	(None, 256)	131328
dropout_109 (Dropout)	(None, 256)	0
dense_133 (Dense)	(None, 64)	16448
dropout_110 (Dropout)	(None, 64)	0
dense_134 (Dense)	(None, 10)	650

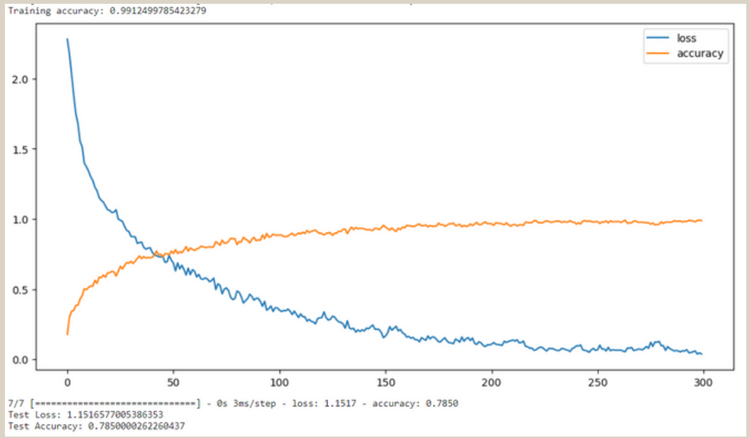
Total params: 178122 (695.79 KB)
Trainable params: 178122 (695.79 KB)
Non-trainable params: 0 (0.00 Byte)

Four layers:

1. Dense layer with 512 neurons and ReLU activation.
2. Dropout layer with a dropout rate of 0.2.
3. Dense layer with 256 neurons and ReLU activation.
4. Dropout layer with a dropout rate of 0.2.
5. Dense layer with 64 neurons and ReLU activation.
6. Dropout layer with a dropout rate of 0.2.
7. Dense layer with 10 neurons (assuming it's a multi-class classification task) and a softmax activation function.

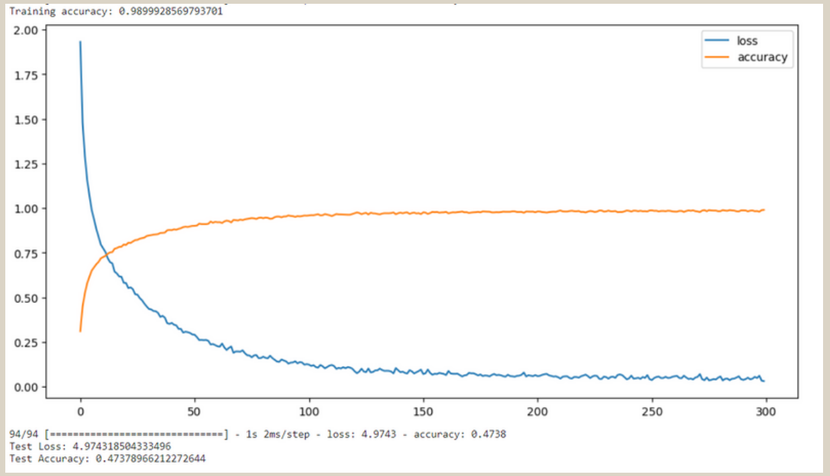
FNN simple

30s dataset



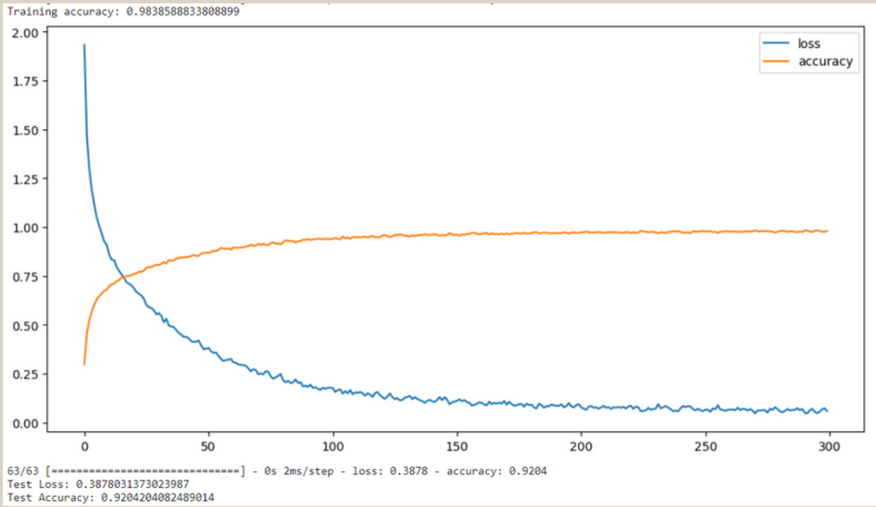
78% accuracy

3s dataset 1st version



47.14% accuracy

3s dataset 2nd version



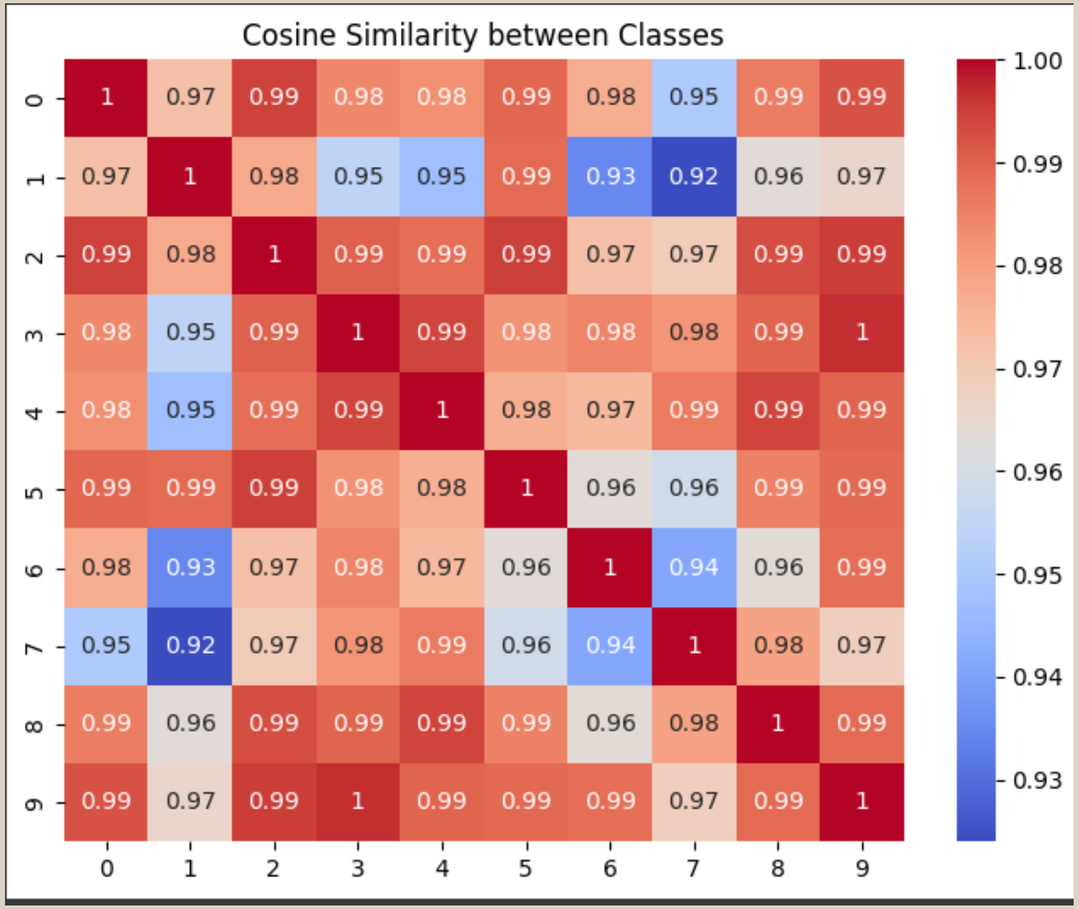
92% accuracy

XGBoost Classifier

- uses decision trees as base learners
- uses a process called pruning to control the size of the decision trees
- based on the principle of gradient boosting, where the algorithm builds a series of weak learners (typically decision trees) sequentially. Each new tree corrects the errors made by the previous ones, with a focus on the examples that were misclassified.

XGBoost Classifier

- 30s dataset - Accuracy : 0.72
- 3s dataset 1st version - Accuracy : 0.72
- 3s dataset 2nd version - 0.90641



Final remark

	precision	recall	f1-score
0	0.71	0.85	0.77
1	0.90	0.90	0.90
2	0.65	0.85	0.74
3	0.85	0.55	0.67
4	0.69	0.90	0.78
5	0.86	0.95	0.90
6	0.93	0.65	0.76
7	0.85	0.85	0.85
8	0.71	0.75	0.73
9	0.64	0.45	0.53

- Low accuracy results for “Metal” and “Classic” music (2 and 9)
- Same problem through all models
- Maybe problem in genre itself, maybe in dataset

Learned skills and take-home messages

- Working with audio data, understanding important features, interpreting data
- Working with different models
- Sometimes more complex model != better one
- Always question others' work
- Be more careful with data split

Thank you for your attention