Group K

Project 01 Report

Abd Alkareem ALJEIRUDI
Stefan MOSER
Markus PROMBERGER

Date: 03.05.2019

Visual analysis

Total class distribution

Features over time

Correlations

Densities

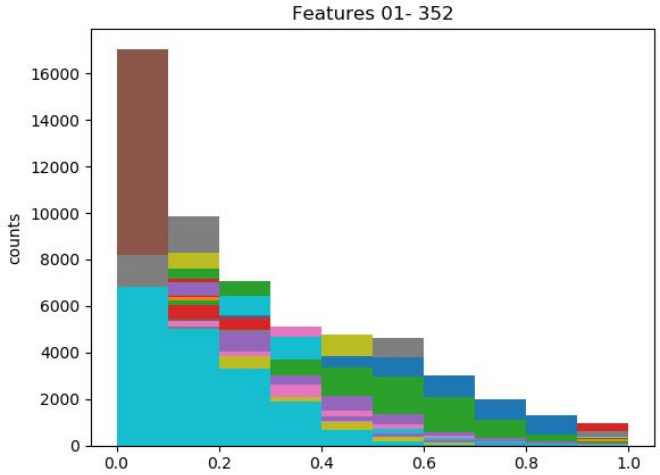
First observations music

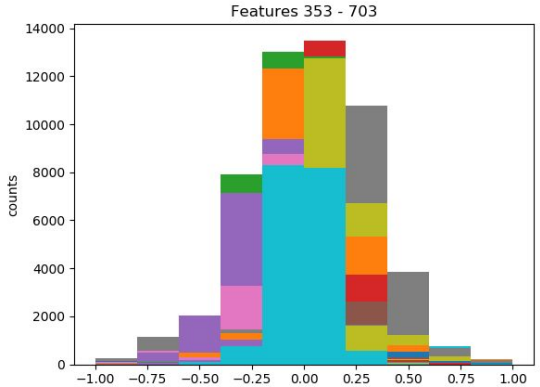
F001 - F352 in range 0 to 1

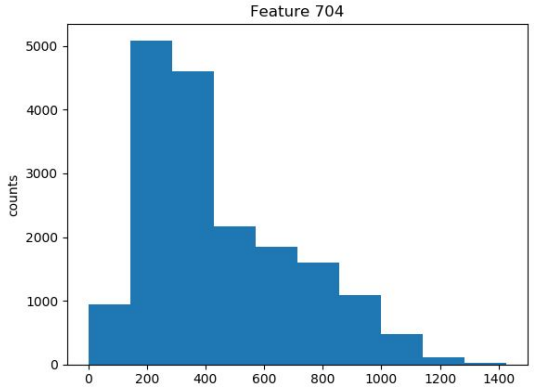
F353 - F703 in range -1 to 1

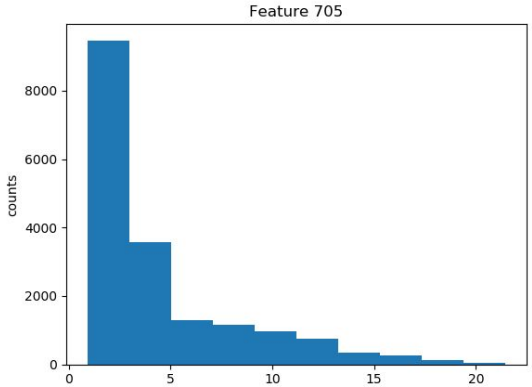
F704 in range 0 to 1500

F705 in range 0 to 25







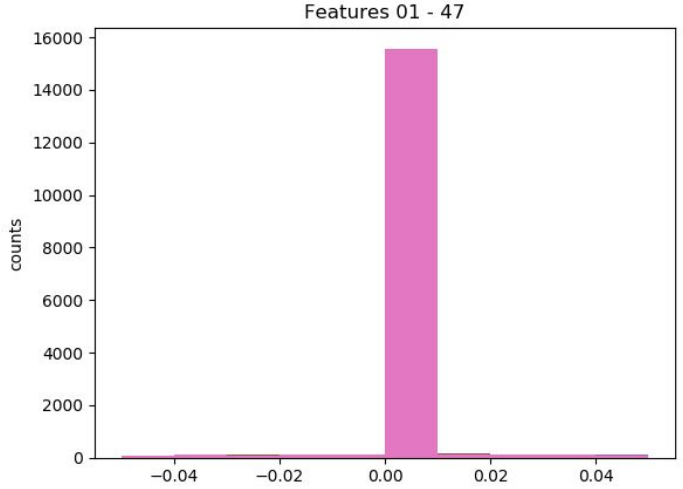


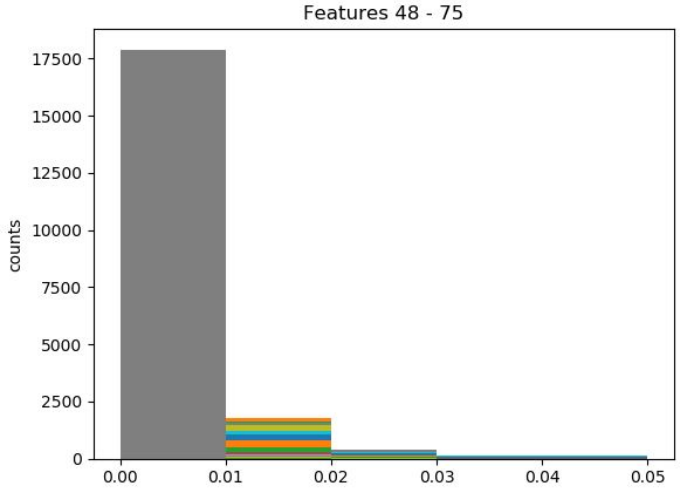
First observations speech

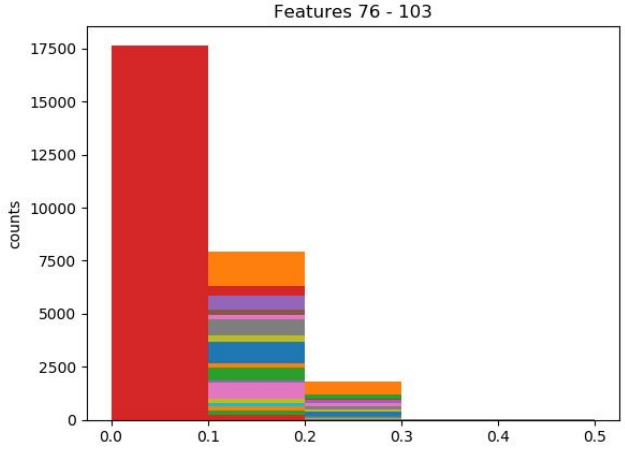
F01 - F47 in range -0.7 to 0.6 (majority 0)

F48 - F75 in range 0 to 0.1

F76 - F103 in range 0 to 0.6



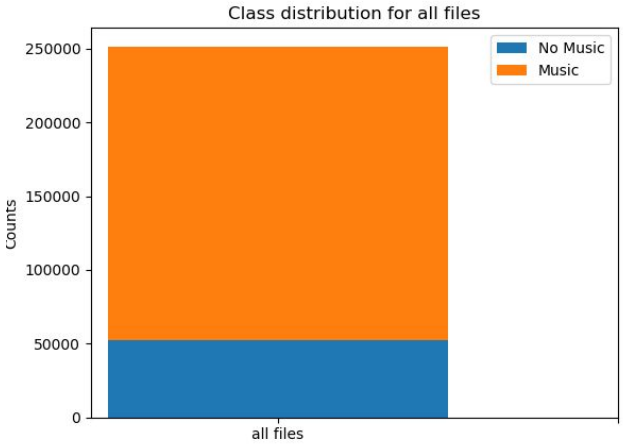


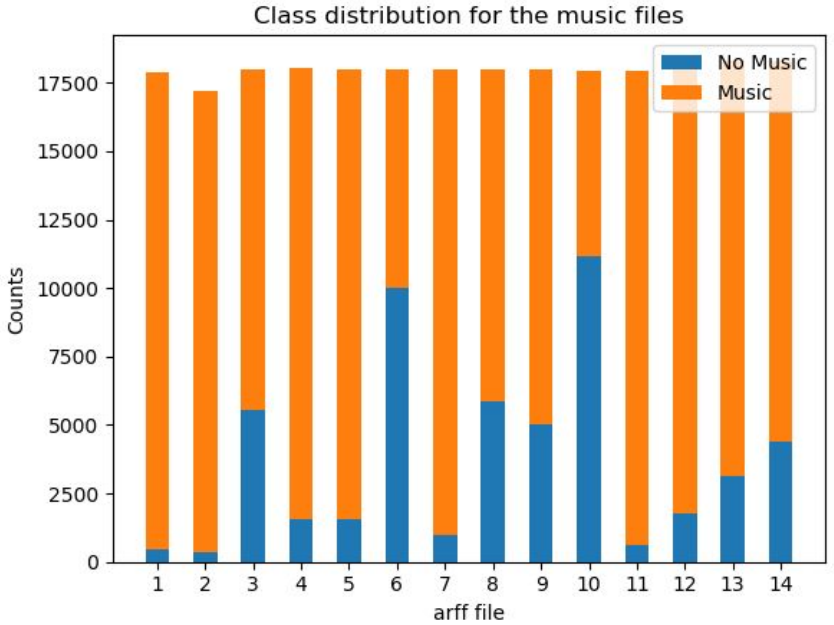


Total class distribution - music

Highly variant distributions in files

In total more “music” than “no music”

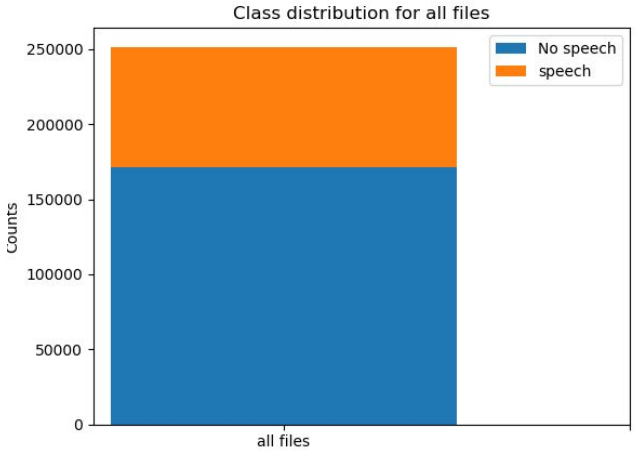


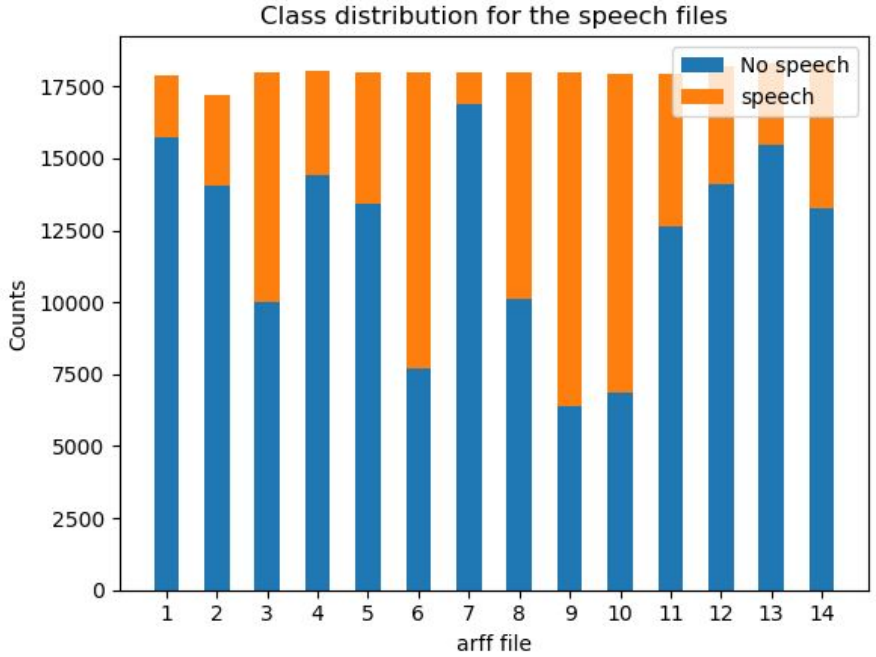


Total class distribution - speech

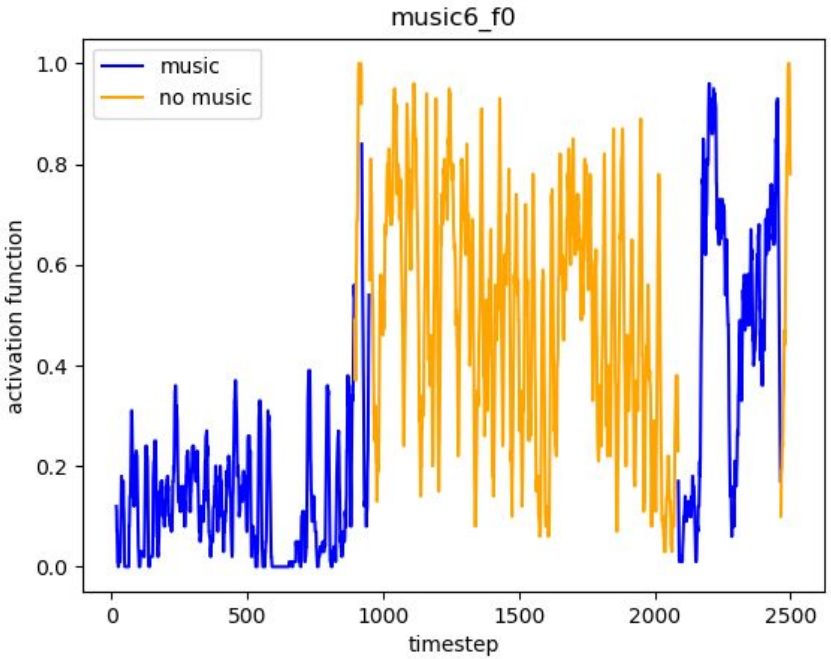
Not as extreme as music

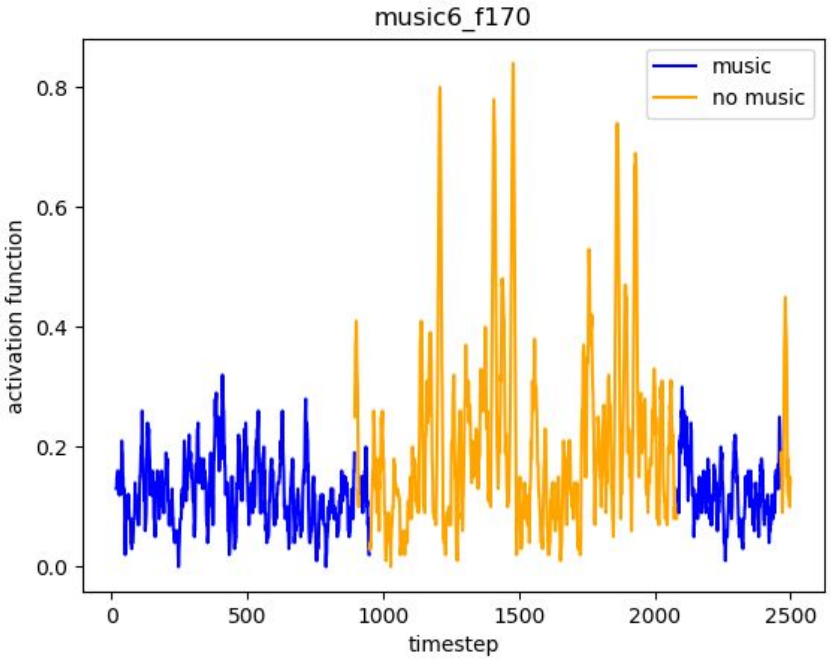
More “no speech” than “speech”



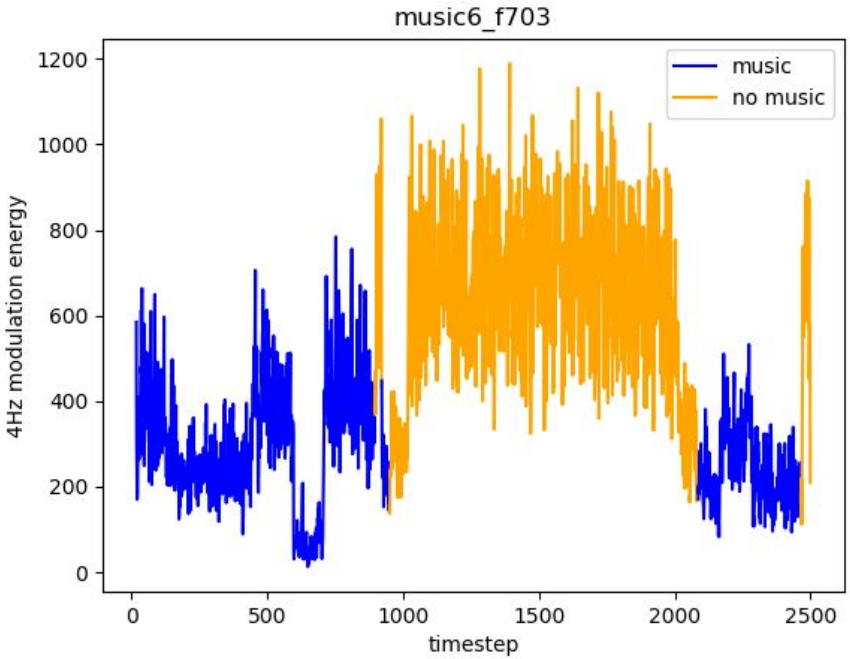


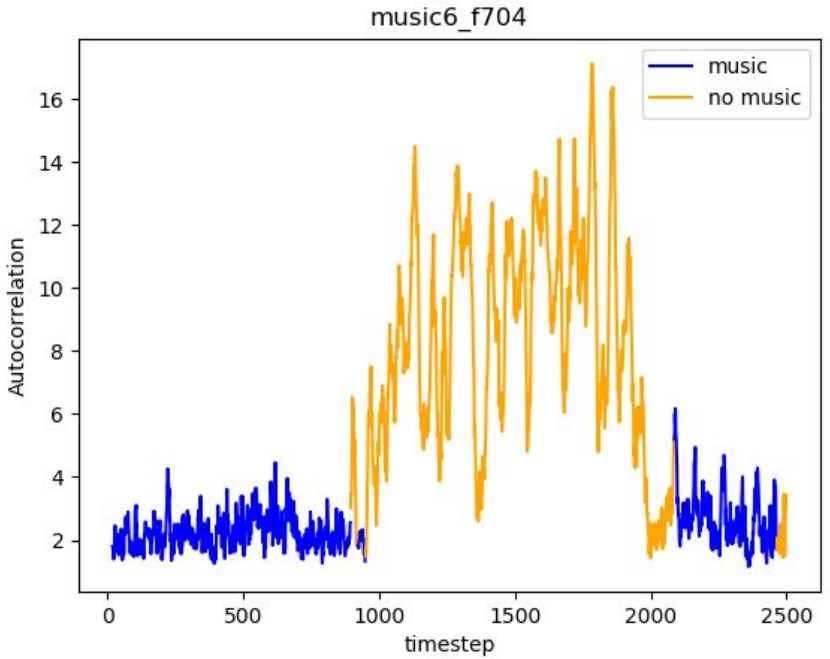
Features over time - music



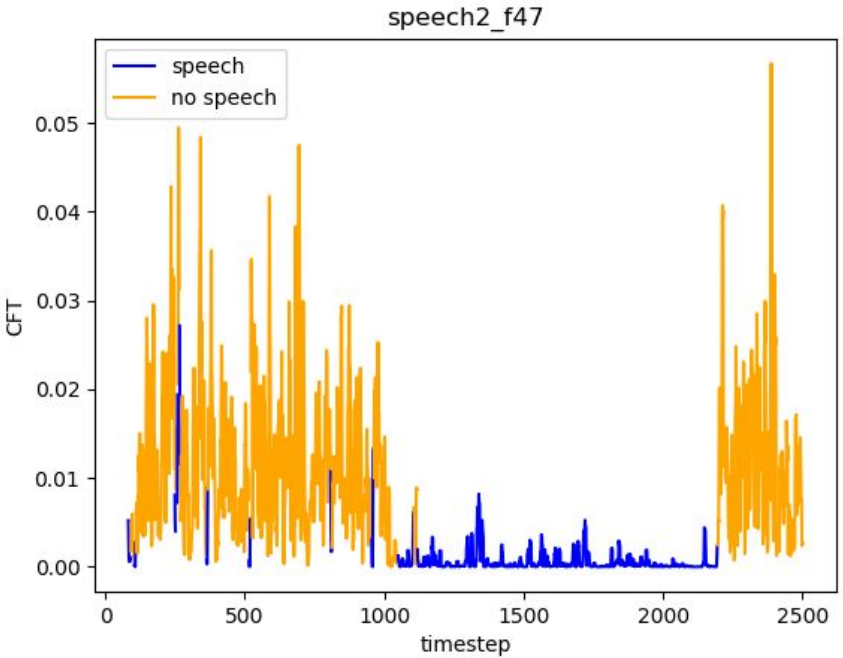


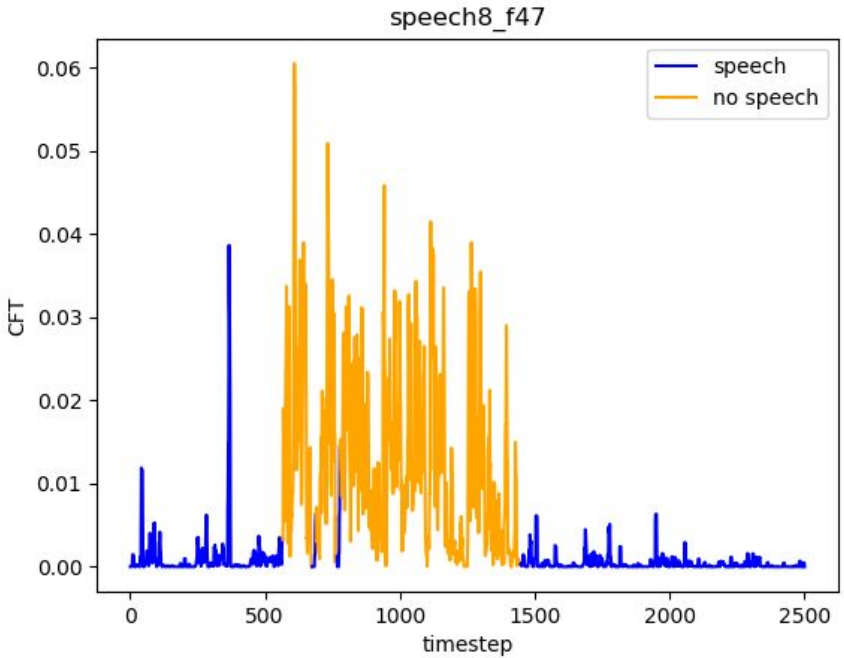
Features over time - music



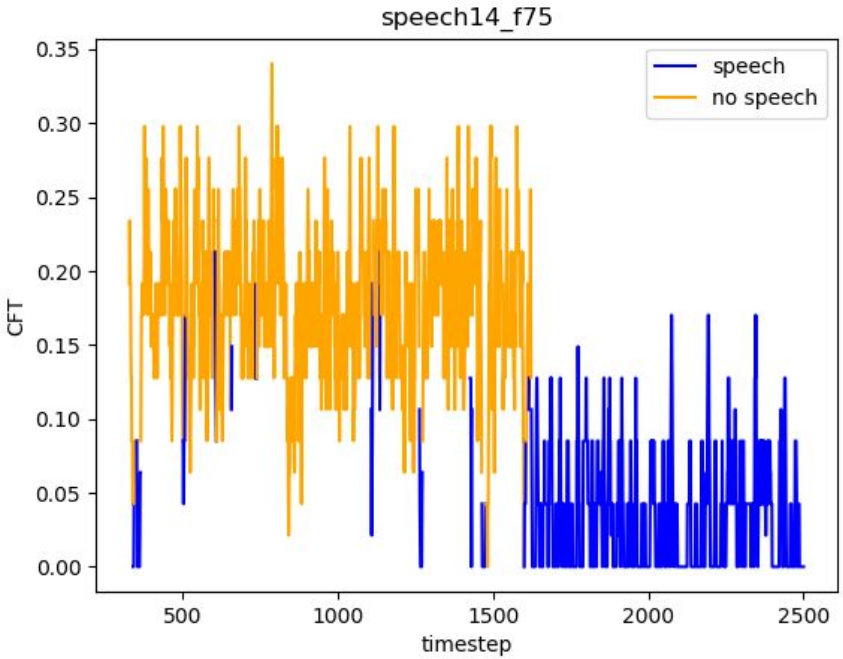


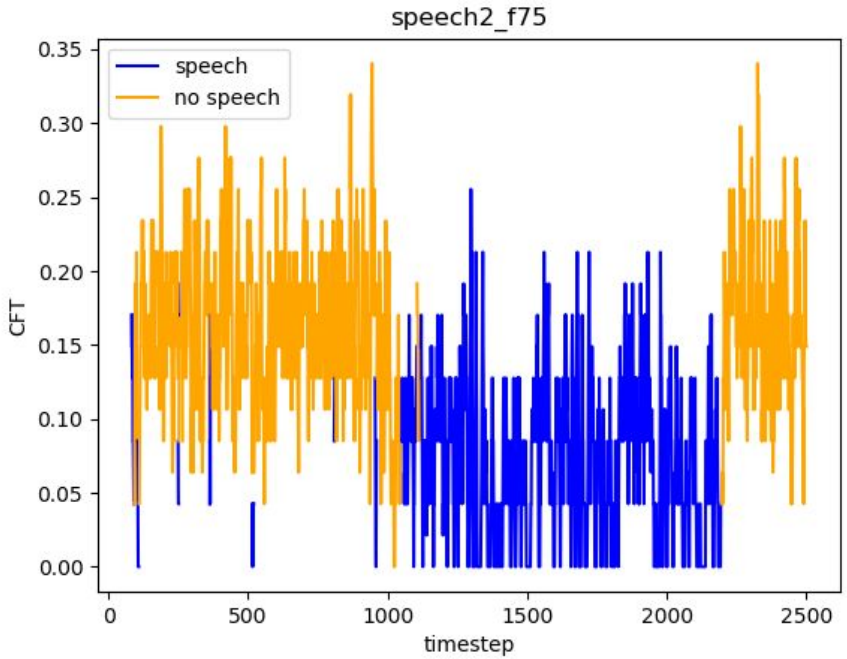
Features over time - speech



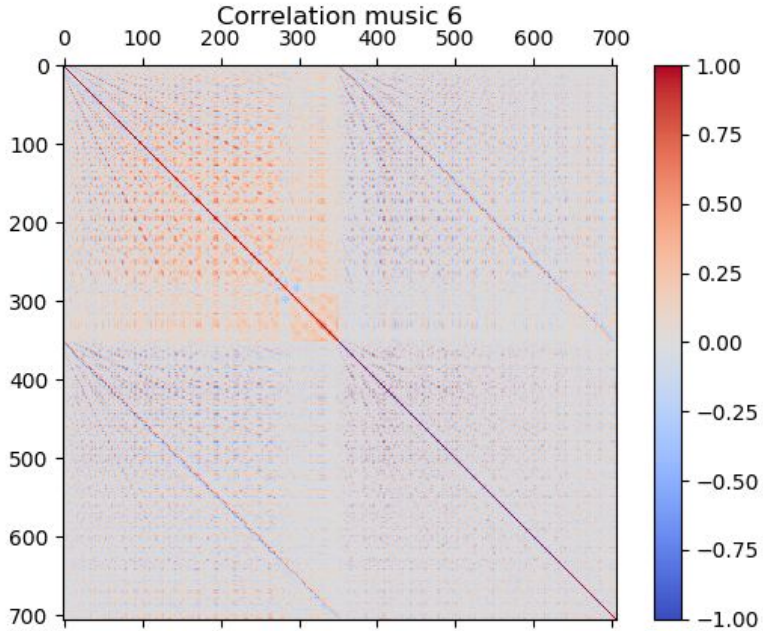


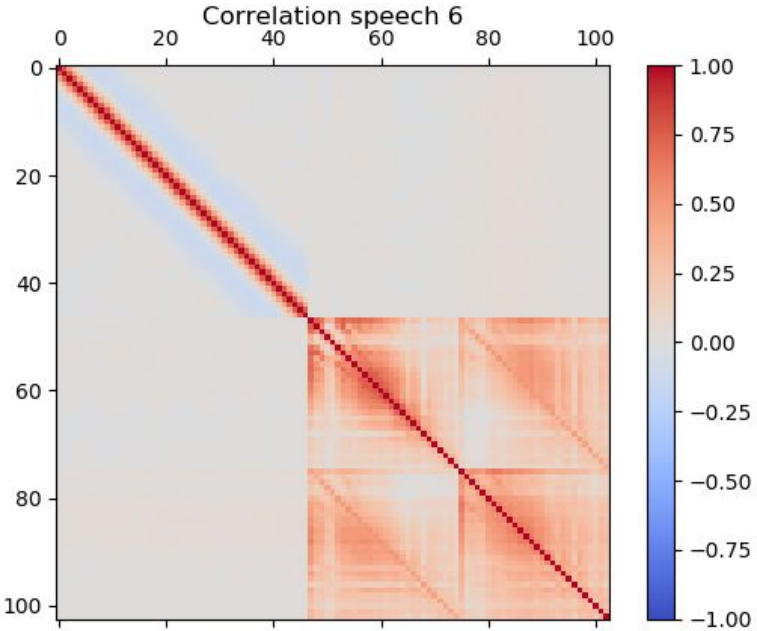
Features over time - speech





Correlation music and speech





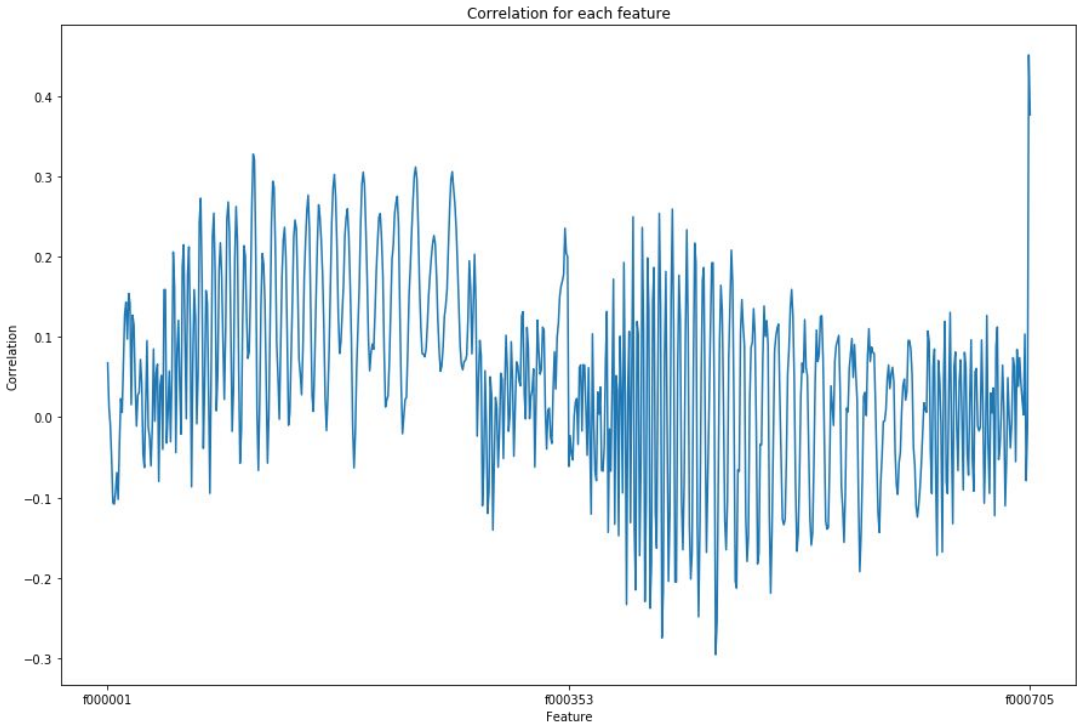
Correlation features - class - music

Pearson-Correlation in music files between features and class

Highly fluctuating

\+ and -

Last 2 features high correlation

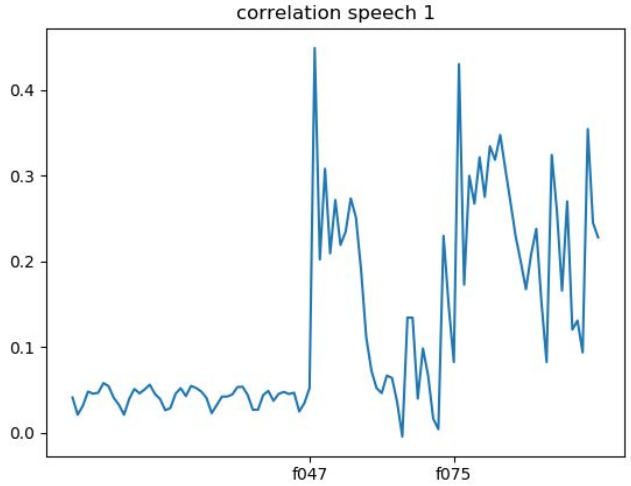


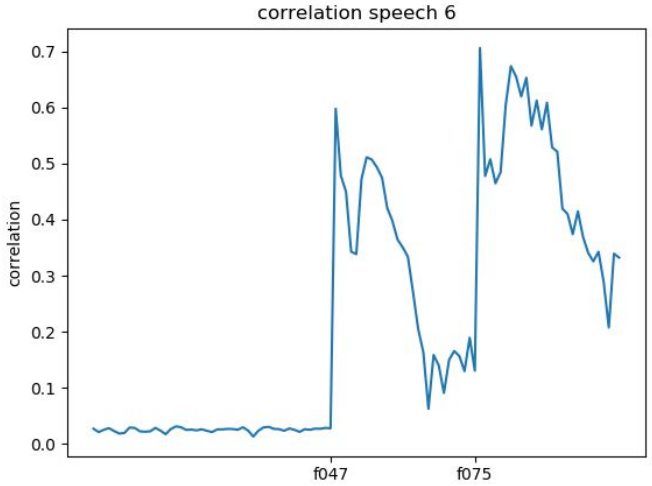
Correlation features - class - speech

Pearson-Correlation in speech files between classes and features

High peak after f047

Low correlation before

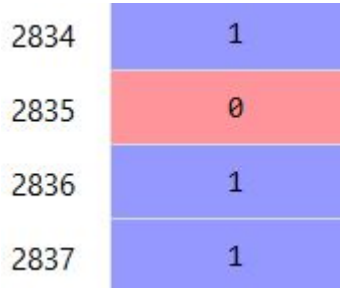




Other observations

Strange class appearances

Probably mislabeling



Classifier Evaluation

5 fold Cross Validation

With Focus on classification accuracy

All data

Using Features with an absolute correlation > 0.2

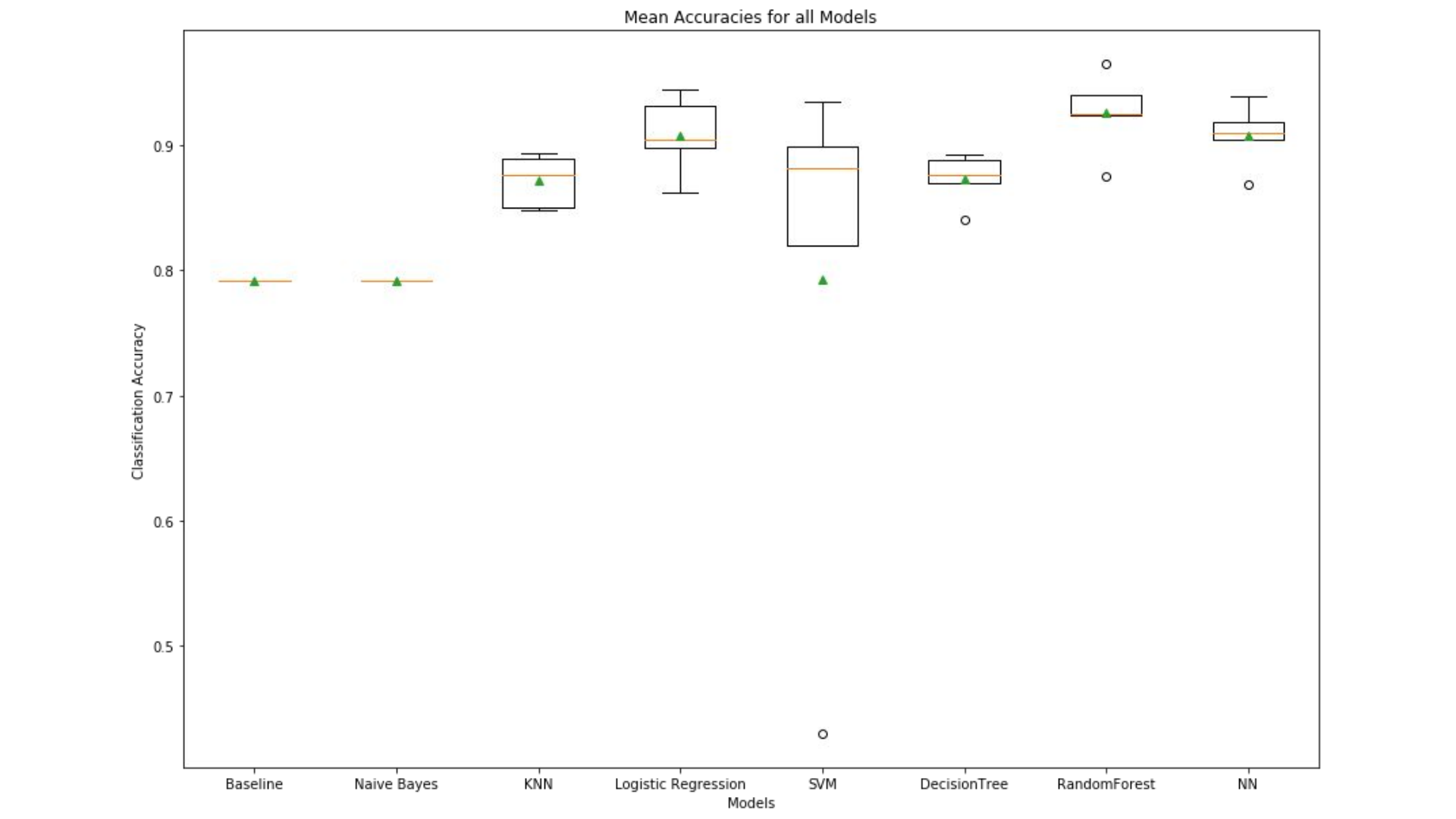
From 705 to 105 features

Classifiers

Classifier	Parameters
DummyClassifier	Most Frequent
Gaussian Naive Bayes	-
Linear Regression	-
K Nearest Neighbors	5 neighbors

Classifiers

Classifier	Parameters
Support Vector Machine	Linear Kernel
Decision Tree	Gini Impurity, best split, min 2 samples for split
Random Forest	100 Trees, Trees as above
Neural Network - Multi-layer Perceptron	1 Hidden Layer with 100 hidden units, Adam Optimizer, Learning Rate 0.001


Mean Accuracies for all Models
Classification Accuracy
0.9
0.8
0.7
0.6
0.5
Baseline
Naive Bayes
KNN
Logistic Regression
SVM
DecisionTree
RandomForest
NN
Models

Best Performance

Random Forest & Neural Network

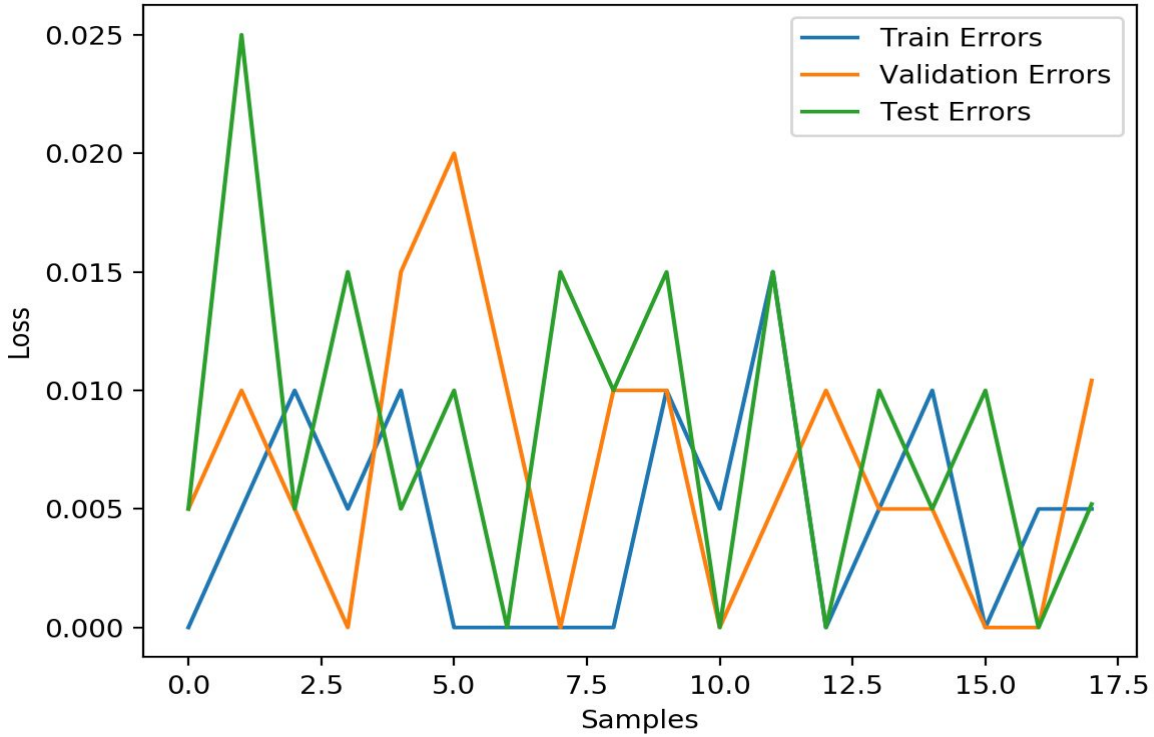
Parameter Evaluation

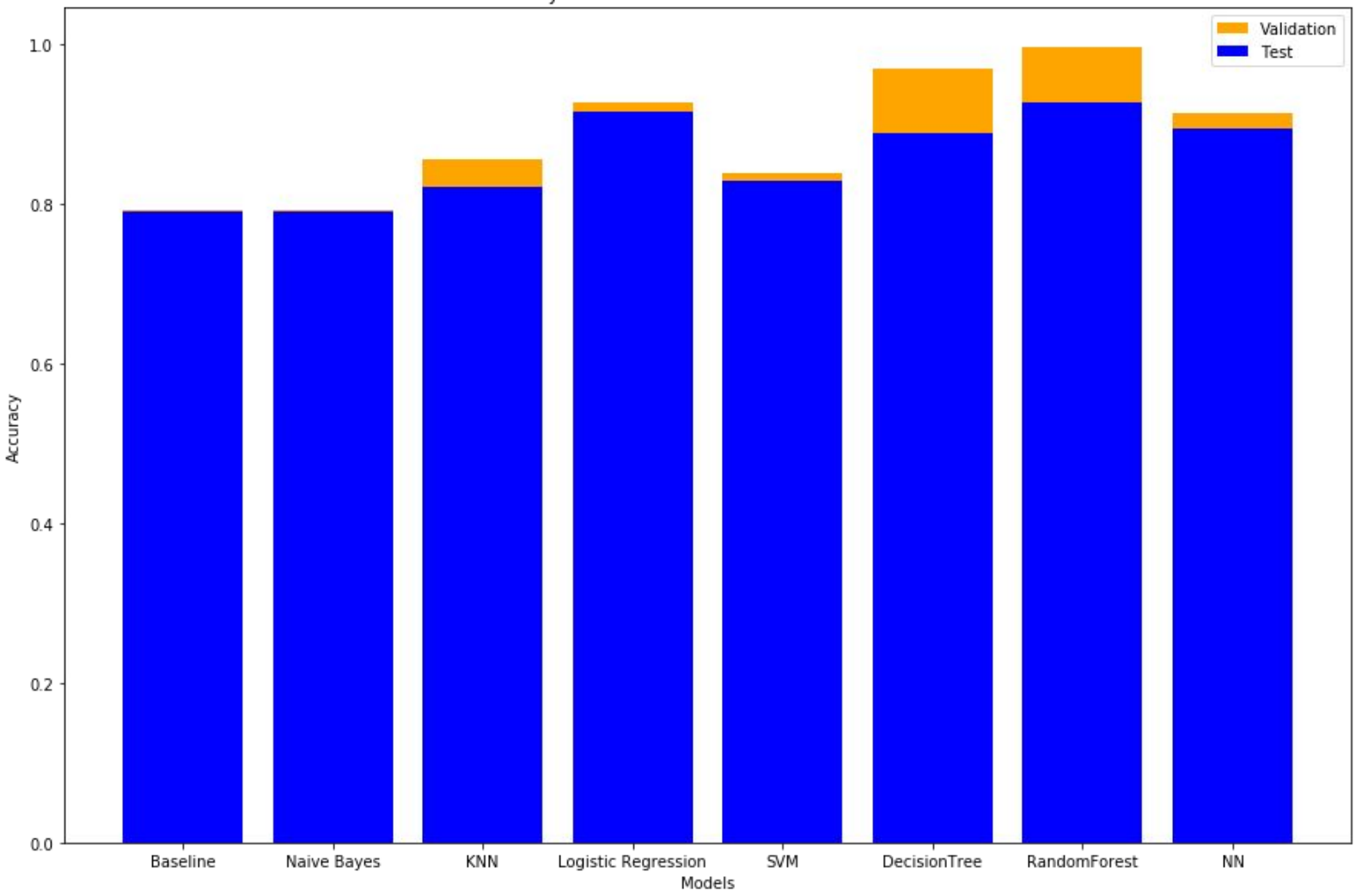
- Using the Bayesian Optimization
- Classification accuracy as an objective function
- 10 exploitation + 5 exploration steps

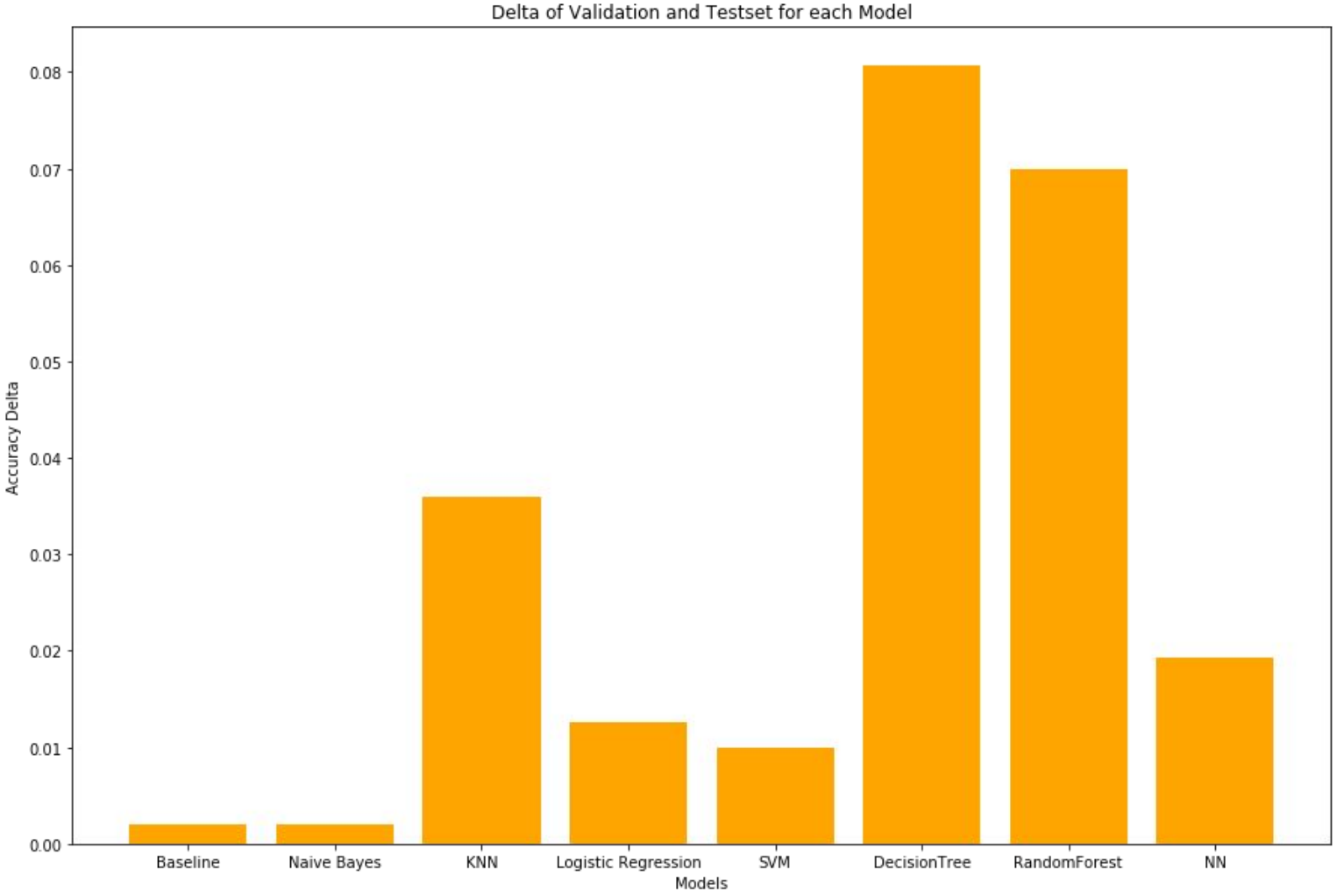
Optimization Results

Model	Best Accuracy	Best Params
NN	0.95967596759676	{'hidden_layer_sizes': 16, 'alpha': 9.263406719097344e-05, 'learning_rate_init': 0.0008804217040183917, 'activation': 'tanh', 'solver': 'adam', 'n_layers': 2}
SVM	0.961856185618562	{'C': 1.0, 'degree': 3, 'gamma': 1.0, 'kernel': 'linear'}
Naive	0.967716771677168	{'none_parameter': 0.05488135039273248}
KNN	0.921893789378938	{'metric': 'euclidean', 'n_neighbors': 13, 'algorithm': 'ball_tree'}
Decision Tree	0.960016201620162	{'min_samples_leaf': 9, 'max_features': 40, 'criterion': 'gini', 'max_depth': 22}
Baseline	0.904032403240324	{'none_parameter': 0.05488135039273248}
RandomForest	0.965716771677168	{'max_features': 37, 'max_depth': 7, 'n_estimators': 62}

Mean-Squared Error

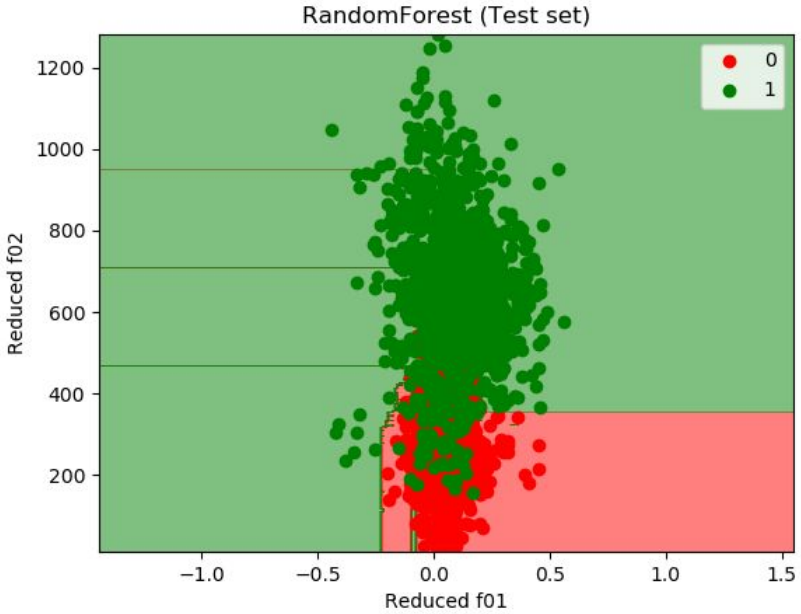


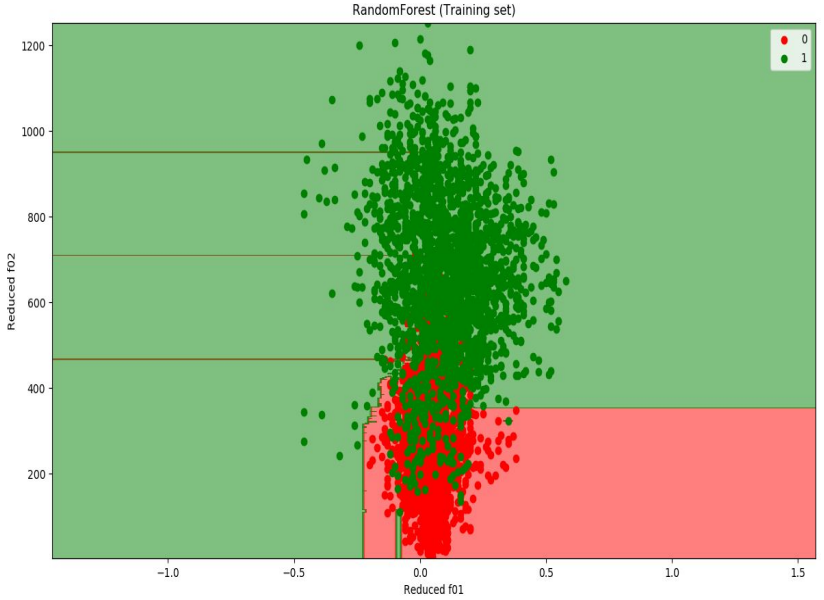

Accuracy of Validation and Testset for each Model
Validation
Test
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Baseline
Naive Bayes
KNN
Logistic Regression
SVM
DecisionTree
RandomForest
NN
Models


Delta of Validation and Testset for each Model
Accuracy Delta
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Baseline
Naive Bayes
KNN
Logistic Regression
SVM
DecisionTree
RandomForest
NN
Models

Overfitting (Train vs. Test Set)?

- RandomForest -> no





Overfitting (Train vs. Test Set)?

- Neural Networks -> no