KU LEUVEN

Faculty of Engineering Technology, Campus Group T

Predicting Genetic Disorders

ML project by Salma Loukili & Nicolas Gutierrez

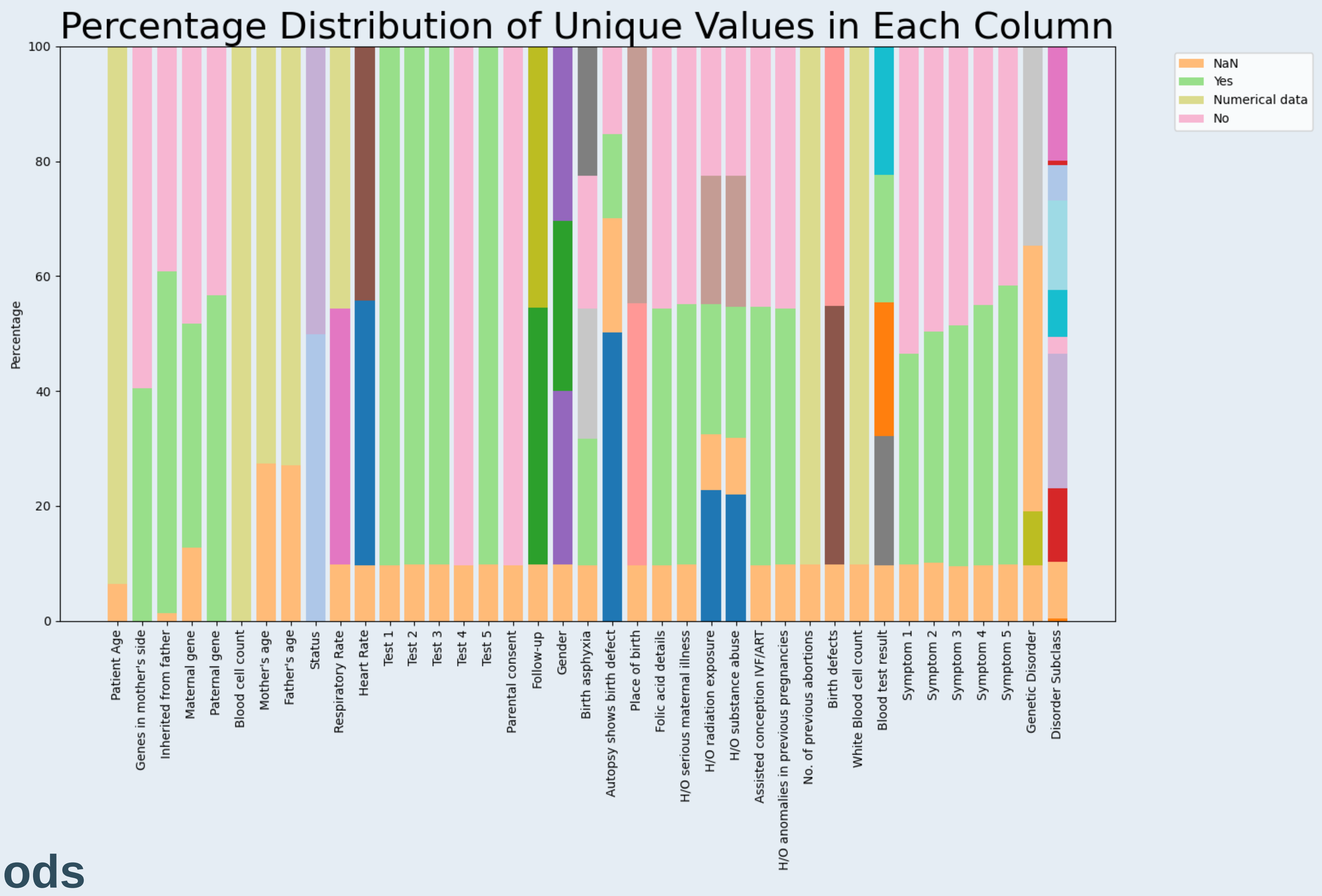
Introduction

In this project, we will be using Machine Learning algorithms to try to predict Genetic Disorders, their subclasses (Disorder Subclass) were used to backfill the missing data.

Previous research on this matter includes the 2022 published paper by the National Library of Medicine titled “Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach”. The research paper uses a hybrid feature selection in which Random Fores and Extra Tree are used to select the best features. Then, XGBoost was the most performing classification model with a 92% α-evaluation score and a 84% macro accuracy score.

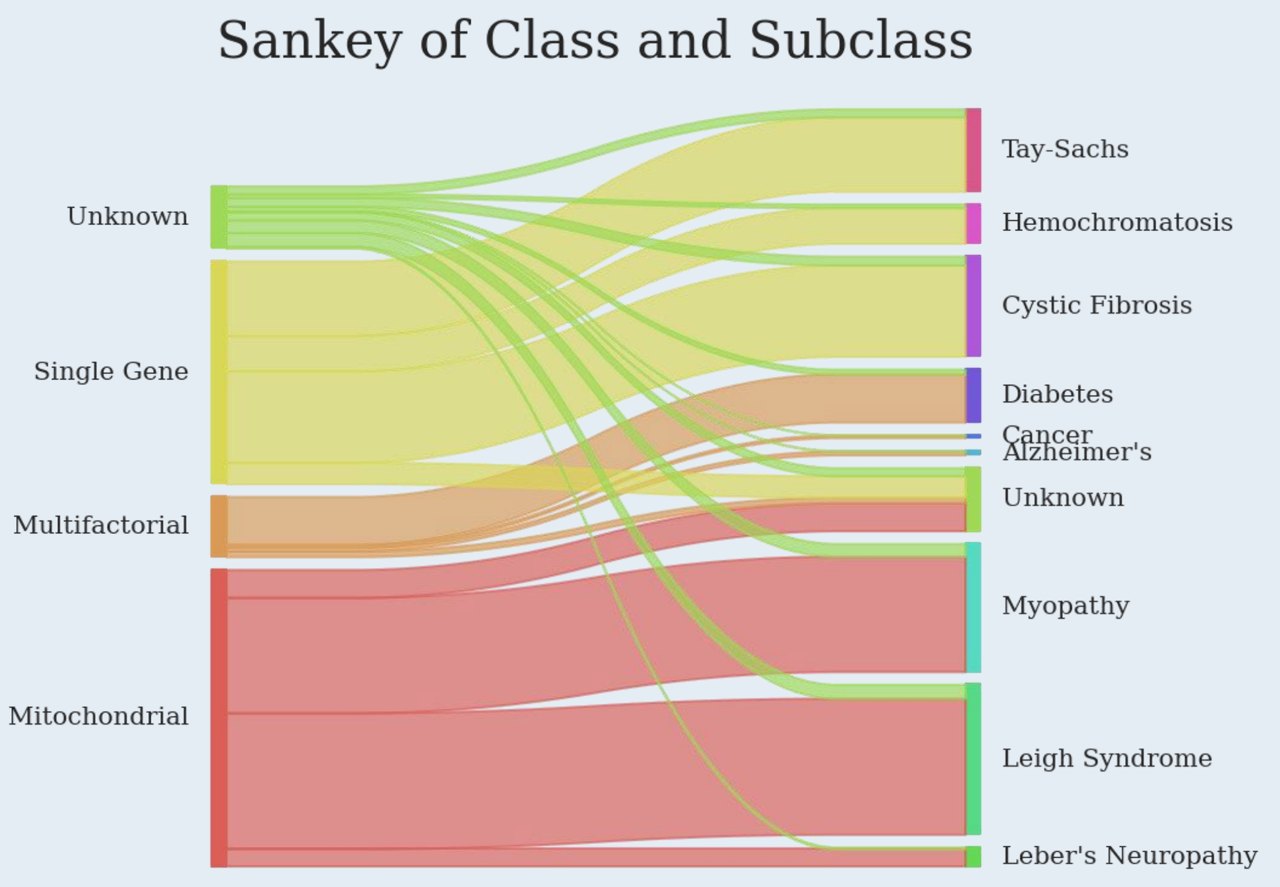
Dataset and Features

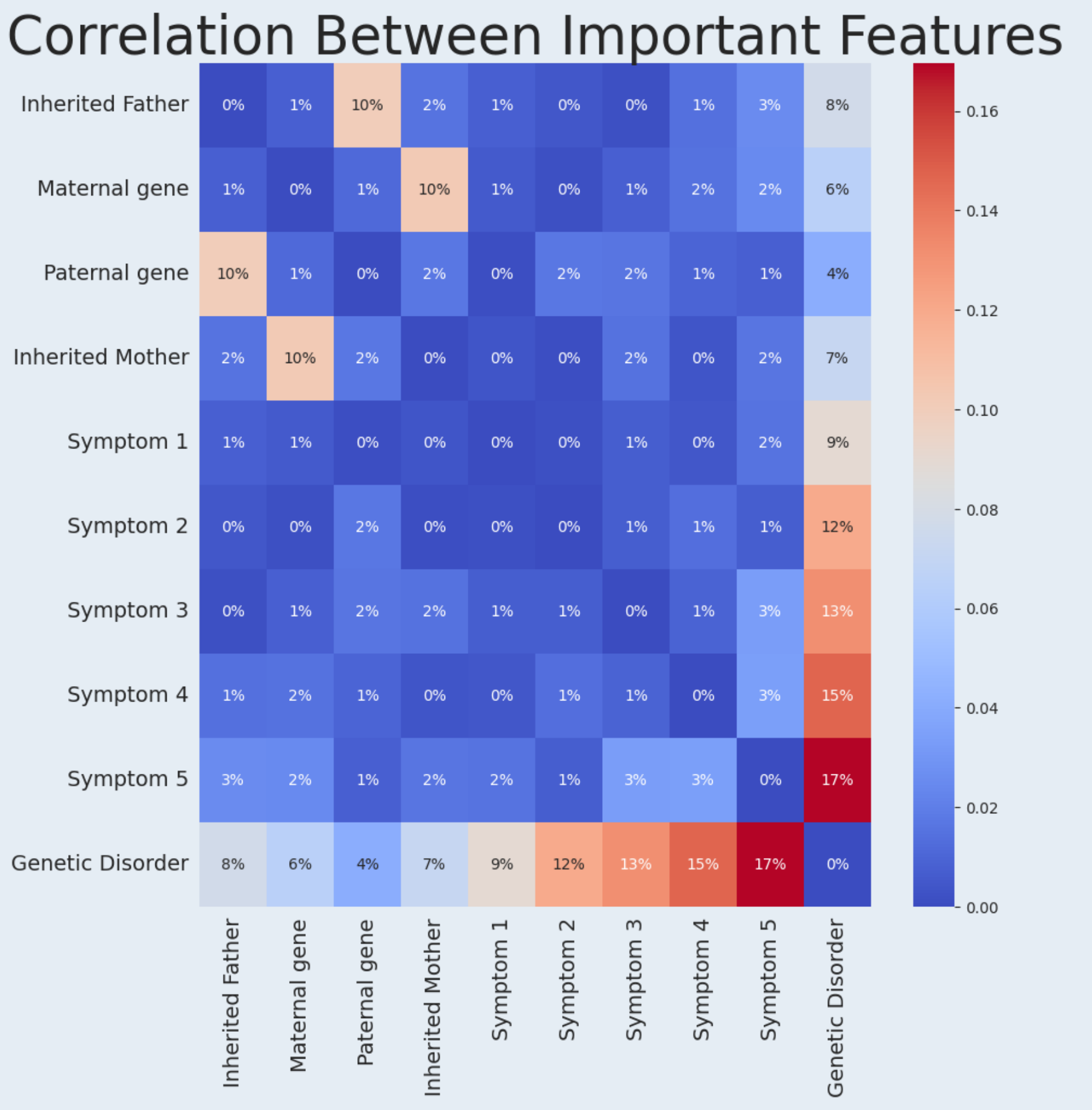
Prior to cleaning and pre-processing, the dataset counts 45 features, 2 of which are Genetic Disorder and Disorder Subclass, these are the two target features.



Methods

The pre-processing of the data included: handling missing values, label/one-hot encoding, scaling, normalizing and feature selection. The heatmap shows the strong correlated features that were kept for training the model.





Neural Network

The first model used is a Keras Sequential Neural Network model (MLP). With a 70/10/20 train/test/validation split, and 4 hidden layers. The hyperparameters were tuned constantly to achieve the best possible results.

Random Forest

The original paper found RF to be well performing when used in conjunction with ETRF feature selection. Nevertheless, this proved to be incredibly difficult to replicate so the most correlated features were selected. They are all categorical-binary, so the processing was very simple. A label was set for each, and they were then scaled.

A RF classifier was set up using hyperparameter tuning. The resulting model was used to predict the correct Genetic Disorder.

Results/Discussion

Random Forest

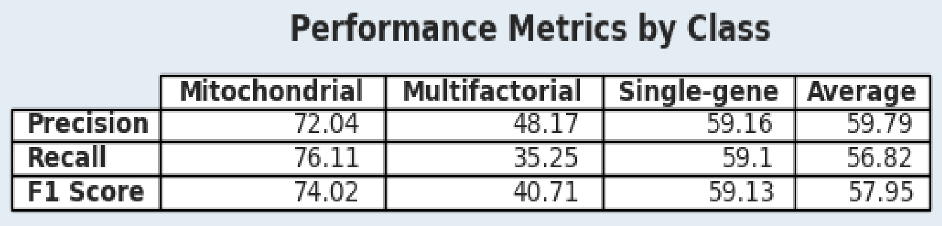
The performance of RF without advanced feature selection techniques proves to be poor. There is not enough data to distinguish between single gene and multifactorial diseases. But when more data was added, the empty values made the dataset too small. KNN imputing was applied but this made the classifier worse. When considering macro precision and F1 score, the model preforms poorly for the smallest class (Multifactorial).

Tensorflow Neural Network

The neural network seems to yield similar results as the RF, especially when it comes to the issue of predicting single gene.

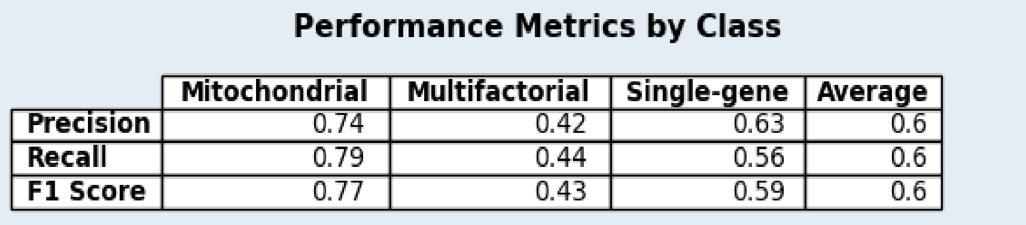
The weighted metrics reach 71 for precision and 65 for the F1 score.

The macro-performance metrics are as follows:



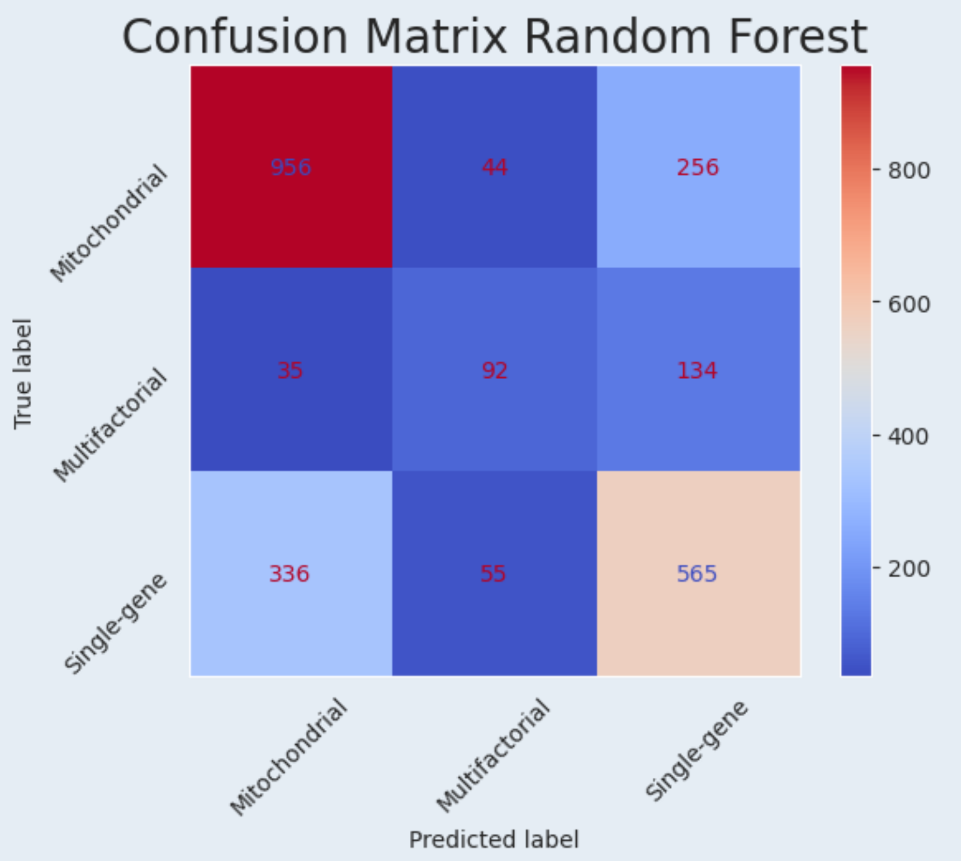
Performance Metrics by Class

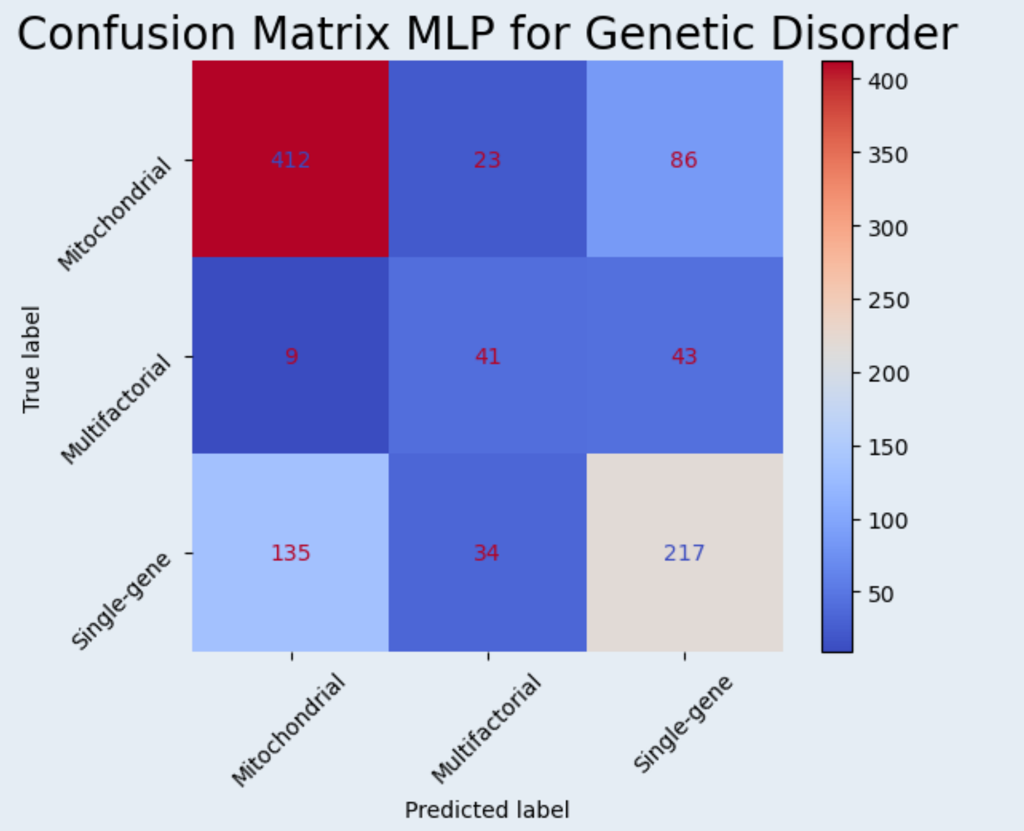
	Mitochondrial	Multifactorial	Single-gene	Average
Precision	72.04	48.17	59.16	59.79
Recall	76.11	35.25	59.1	56.82
F1 Score	74.02	40.71	59.13	57.95

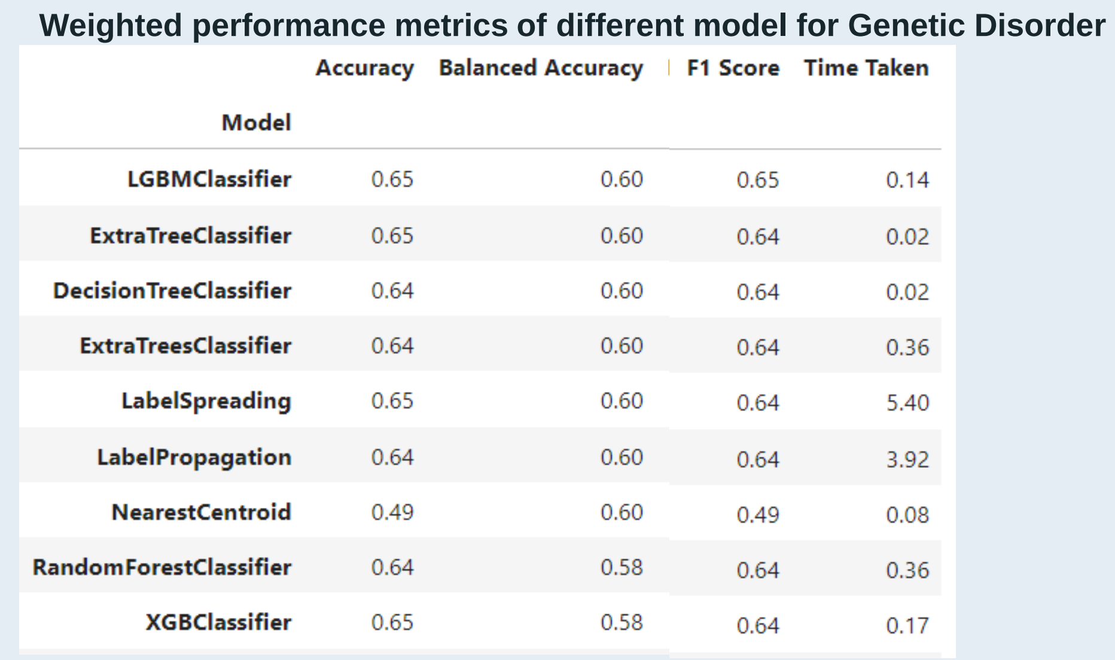


Performance Metrics by Class

	Mitochondrial	Multifactorial	Single-gene	Average
Precision	0.74	0.42	0.63	0.6
Recall	0.79	0.44	0.56	0.6
F1 Score	0.77	0.43	0.59	0.6

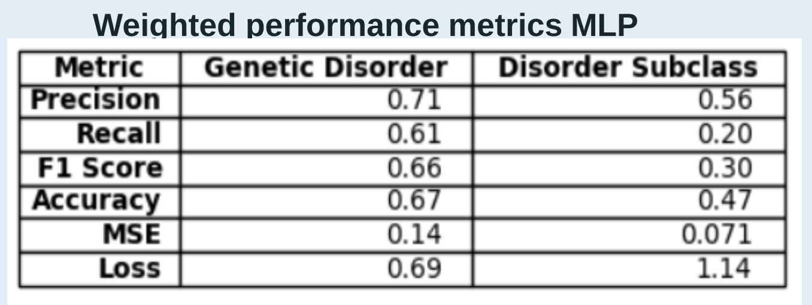






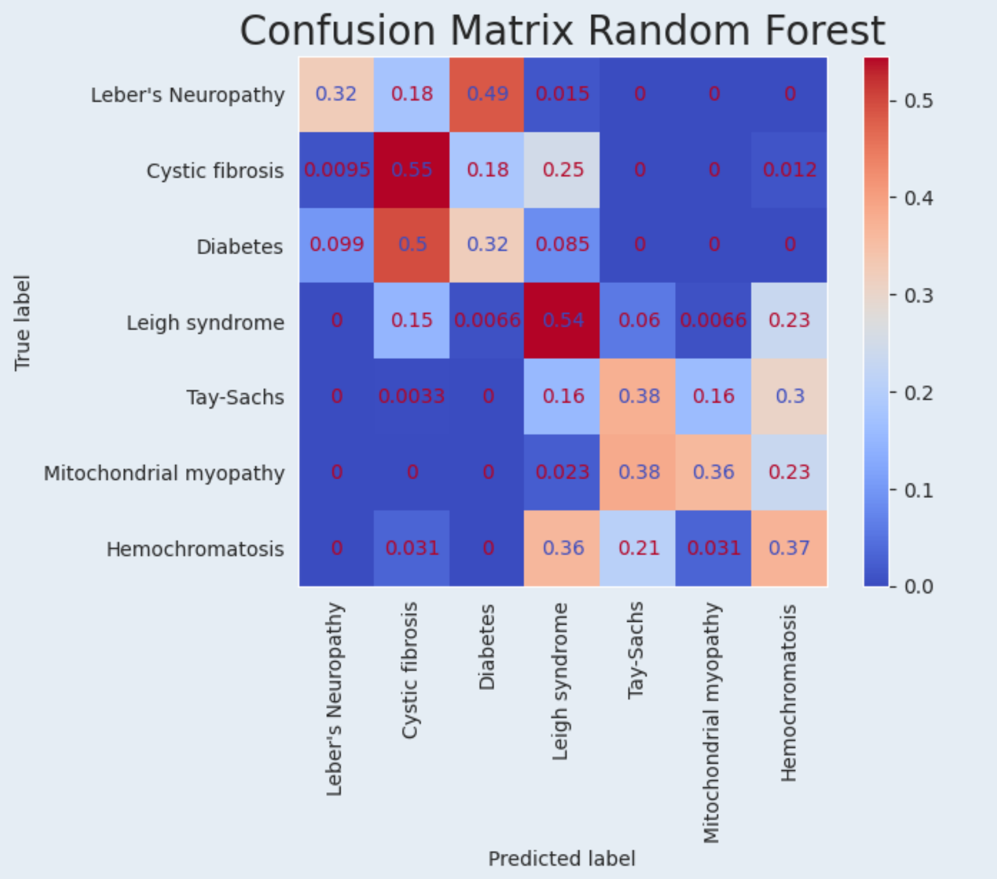
Weighted performance metrics of different model for Genetic Disorder

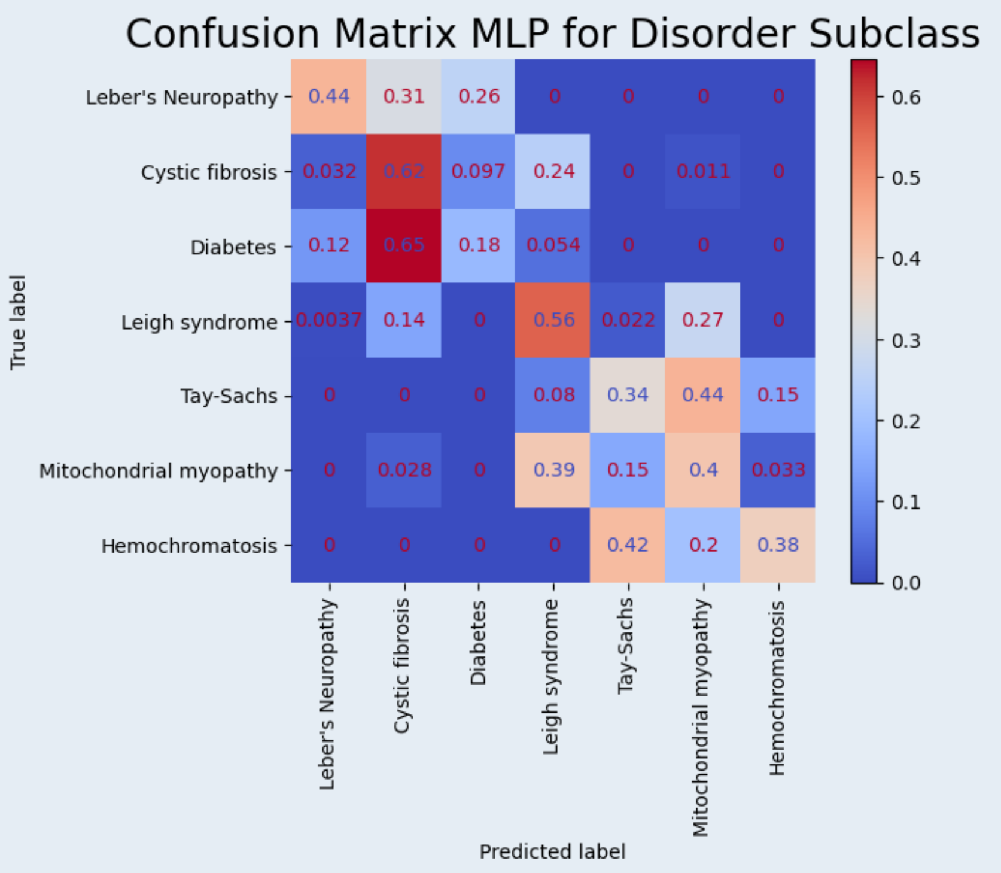
Model	Accuracy	Balanced Accuracy	F1 Score	Time Taken
LGBMClassifier	0.65	0.60	0.65	0.14
ExtraTreeClassifier	0.65	0.60	0.64	0.02
DecisionTreeClassifier	0.64	0.60	0.64	0.02
ExtraTreesClassifier	0.64	0.60	0.64	0.36
LabelSpreading	0.65	0.60	0.64	5.40
LabelPropagation	0.64	0.60	0.64	3.92
NearestCentroid	0.49	0.60	0.49	0.08
RandomForestClassifier	0.64	0.58	0.64	0.36
XGBClassifier	0.65	0.58	0.64	0.17



Weighted performance metrics MLP

Metric	Genetic Disorder	Disorder Subclass
Precision	0.71	0.56
Recall	0.61	0.20
F1 Score	0.66	0.30
Accuracy	0.67	0.47
MSE	0.14	0.071
Loss	0.69	1.14





Conclusion/Future Work

- Dataset with many irrelevant or empty features, complicating analysis.
- Lack of detail on ETRF construction.
- Dimensionality reduction and better feature selection might enhance results.
- Need to identify discriminant features for Single-Gene vs. Multifactorial diseases.
- SMOTE reduced overfitting but decreased classifier performance.
- Models overfit on certain classes and underpredict Multifactorial diseases.
- Further analysis required to distinguish Multifactorial diseases.