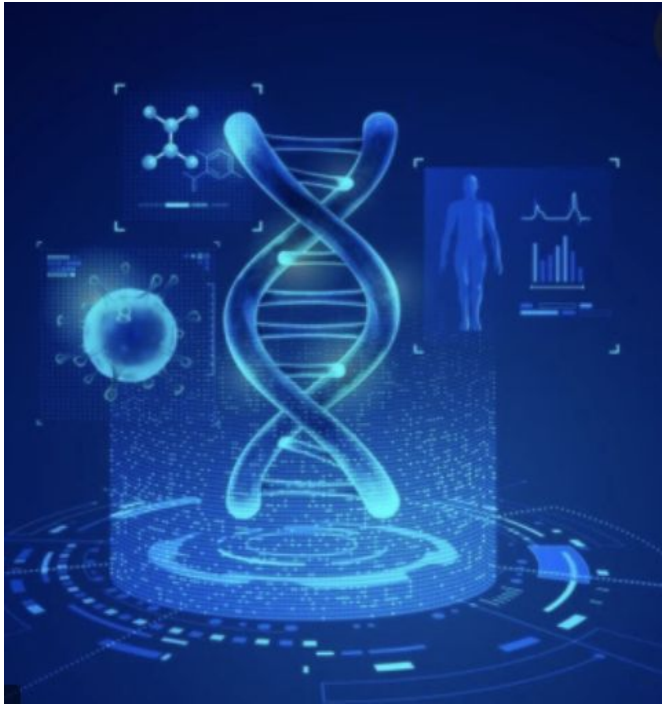
Genomes and Genetics

Brooke Hanson

The Problem

- An exponential increase in population has lead to an increase in genetic mutations and hereditary illnesses
- Understanding what causes these illnesses is essential to preventing and promoting healthy births



Can these disorders be predicted?

Who might care?

- Hereditary Illnesses affect individuals of all walks of life
- Individuals who are hoping to conceive
- Family members who are involved in caretaking
- Doctors who treat any of these patients

Features That May Affect Hereditary Illnesses

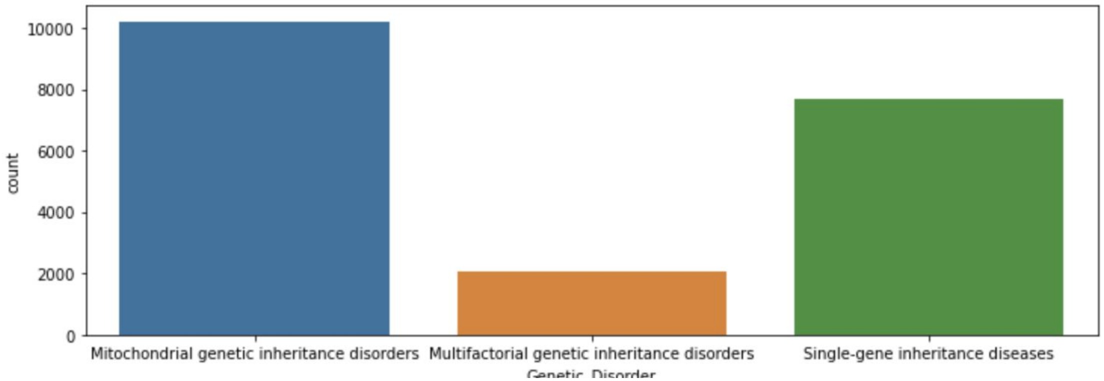
- The features in the data that may have the most affect are
  - Genes in mother’s side
  - Maternal Gene
  - Paternal Gene
  - Blood Test Result

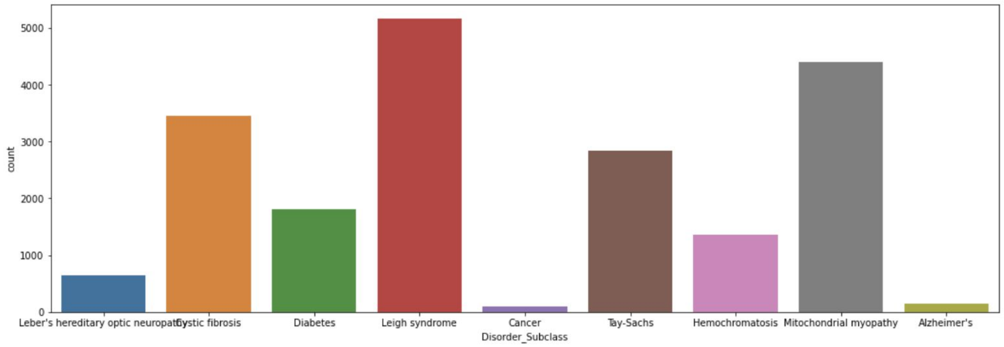
Data Information

- Dataset from Kaggle.com
- 18,047 rows of data
- 29 columns for first model
- 30 columns for second model
    - Patient_Age , Genes_in_mothers_side , Inherited_from_father ,Maternal_gene, Paternal_gene, Blood_cell_count ,Mothers_age, Fathers_age, Respiratory_Rate, Heart_Rate, Parental_consent, Follow_up, Gender, Birth_asphyxia, Autopsy_shows_birth_defect, Folic_acid_details, HO_serious_maternal_illness, HO_radiation_exposure, HO_substance_abuse, Assisted_conception_IVFART, History_of_anomalies_in_previous_pregnancies, No_of_previous_abortion, Birth_defects, White_Blood_cell_count, Blood_test_result, Symptom_1, Symptom_2, Symptom_3, Symptom_4, Symptom_5, Genetic_Disorder, Disorder_Subclass
    -

Exploratory Data Analysis

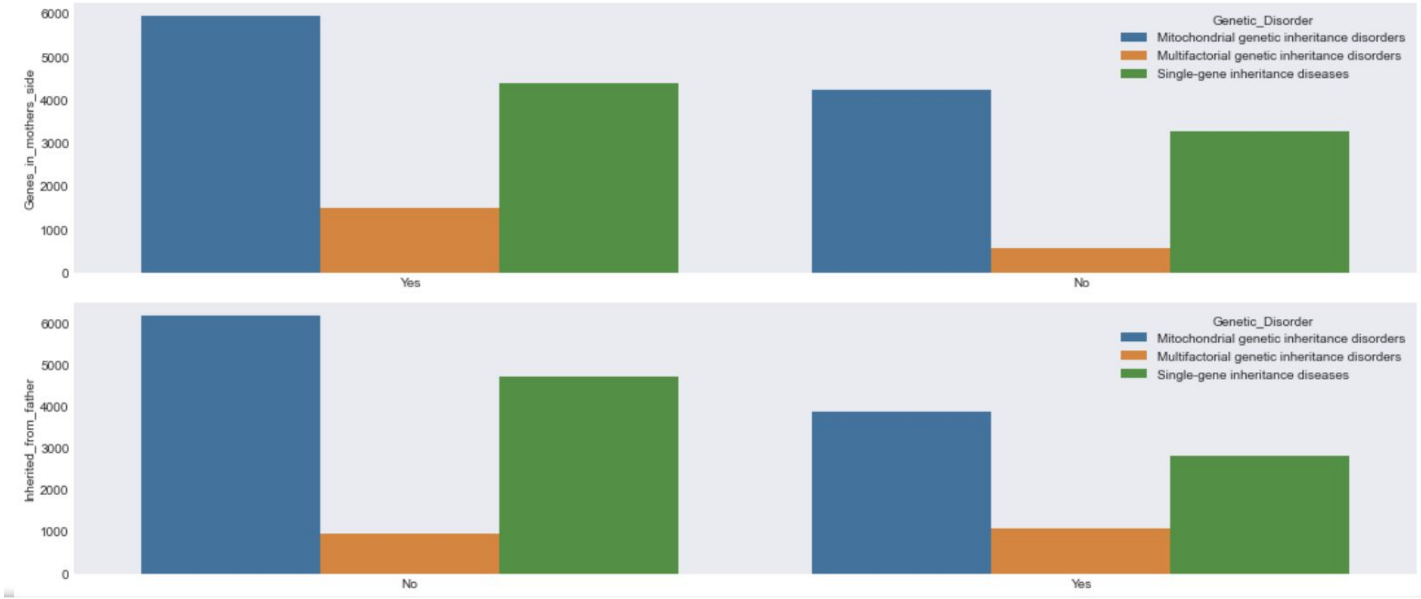
- Started by plotting the counts of each type of Genetic Disorder and each Disorder Subclass





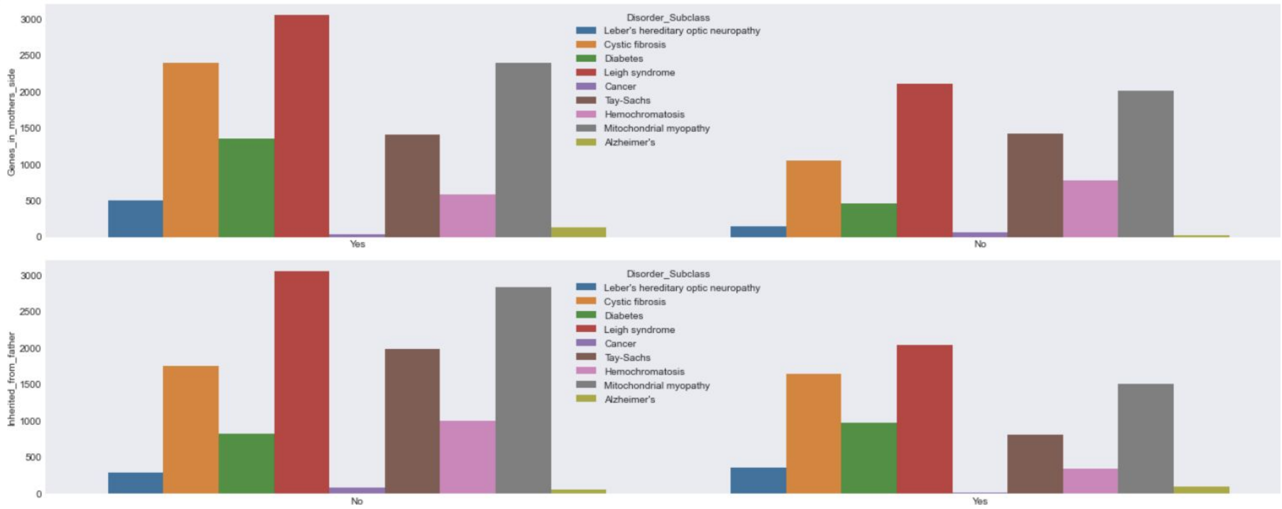
Exploratory Data Analysis

- A first look at how categorical variables interact with Genetic Disorder

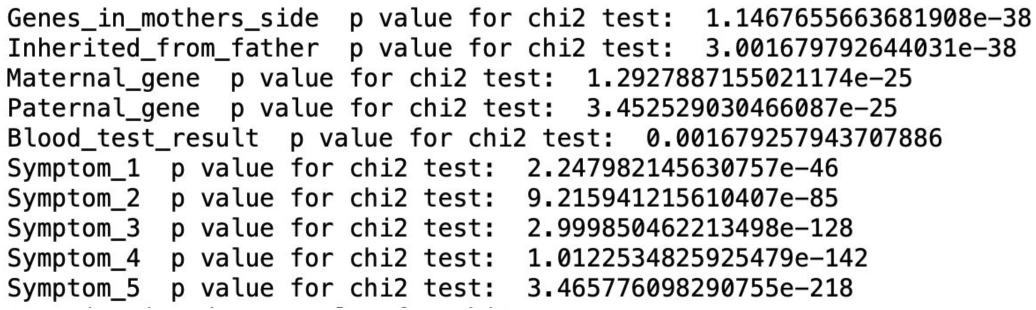


Exploratory Data Analysis

- A look at how categorical variables affect the Disorder Subclass

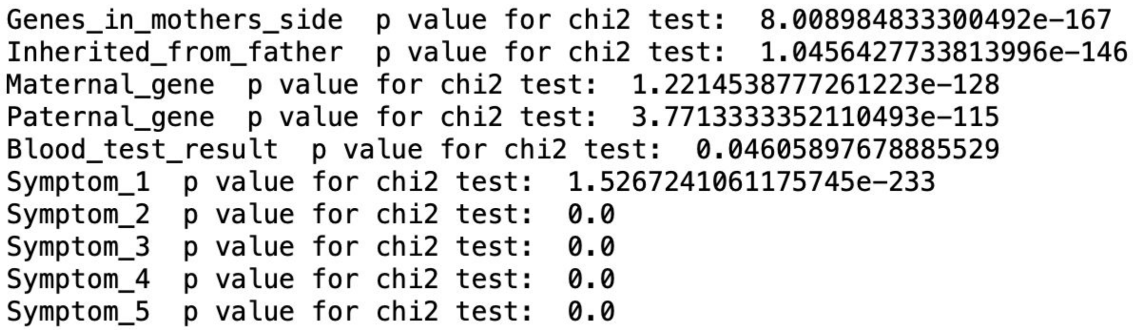


Exploratory Data Analysis



```
Genes_in_mothers_side  p value for chi2 test:  1.1467655663681908e-38
Inherited_from_father  p value for chi2 test:  3.001679792644031e-38
Maternal_gene  p value for chi2 test:  1.2927887155021174e-25
Paternal_gene  p value for chi2 test:  3.452529030466087e-25
Blood_test_result  p value for chi2 test:  0.0016792579437078 86
Symptom_1  p value for chi2 test:  2.247982145630757e-46
Symptom_2  p value for chi2 test:  9.215941215610407e-85
Symptom_3  p value for chi2 test:  2.999850462213498e-128
Symptom_4  p value for chi2 test:  1.0122534825925479e-142
Symptom_5  p value for chi2 test:  3.465776098290755e-218
```

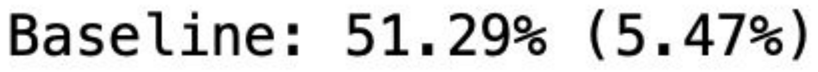
- Series of chi2 tests to identify independence of variables
- Left - Genetic Disorder
- Right - Disorder Subclass



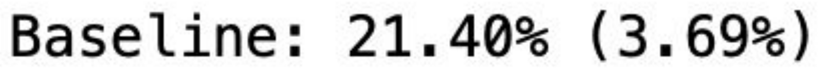
```
Genes_in_mothers_side  p value for chi2 test:  8.008984833300492e-167
Inherited_from_father  p value for chi2 test:  1.0456427733813996e-146
Maternal_gene  p value for chi2 test:  1.2214538777261223e-128
Paternal_gene  p value for chi2 test:  3.7713333352110493e-115
Blood_test_result  p value for chi2 test:  0.04605897678885529
Symptom_1  p value for chi2 test:  1.5267241061175745e-233
Symptom_2  p value for chi2 test:  0.0
Symptom_3  p value for chi2 test:  0.0
Symptom_4  p value for chi2 test:  0.0
Symptom_5  p value for chi2 test:  0.0
```

Modeling

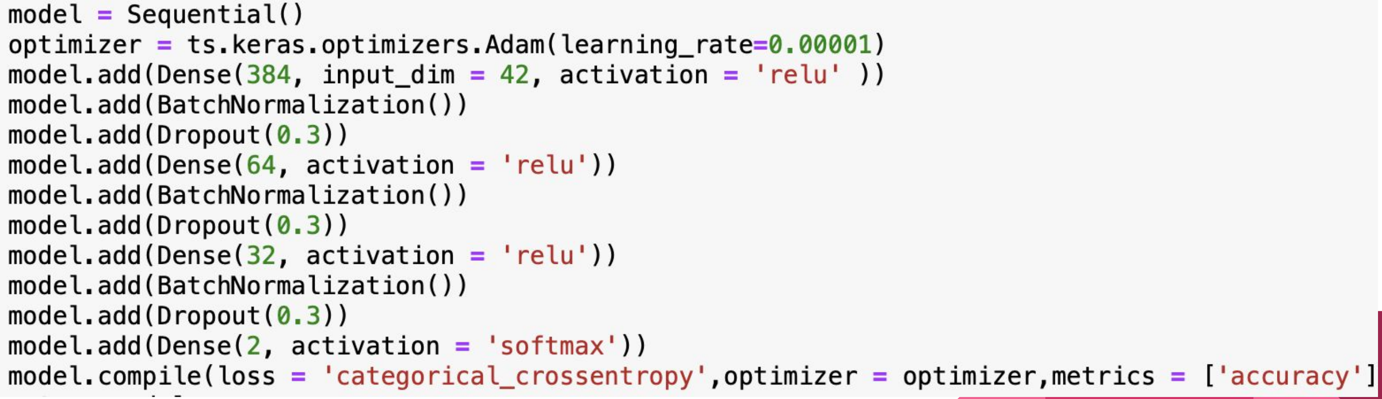
- Model used
- Made changes to optimizers to tune model
- Included BatchNormalization(), Dropout(), and used soft max for final activation



```
Baseline: 51.29% (5.47%)
```

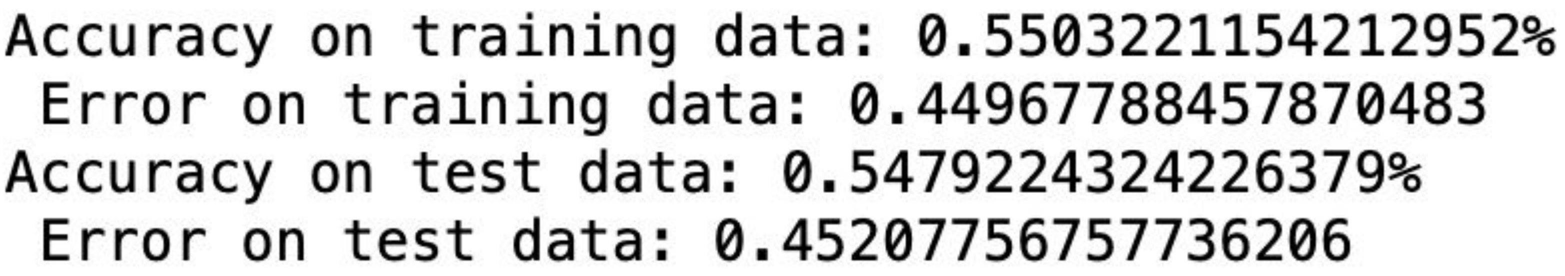


```
Baseline: 21.40% (3.69%)
```

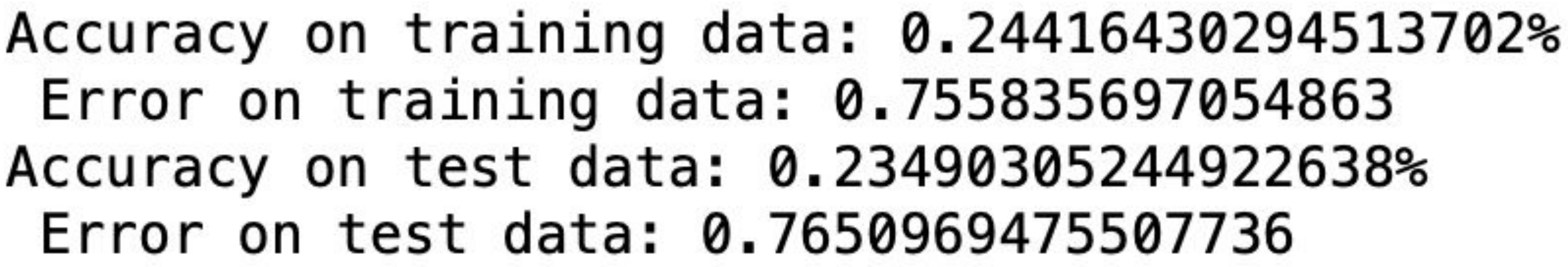


```
model = Sequential()
optimizer = ts.keras.optimizers.Adam(learning_rate=0.00001)
model.add(Dense(384, input_dim = 42, activation = 'relu' ))
model.add(BatchNormalization())
model.add(Dropout(0.3))
model.add(Dense(64, activation = 'relu'))
model.add(BatchNormalization())
model.add(Dropout(0.3))
model.add(Dense(32, activation = 'relu'))
model.add(BatchNormalization())
model.add(Dropout(0.3))
model.add(Dense(2, activation = 'softmax'))
model.compile(loss = 'categorical_crossentropy',optimizer = optimizer,metrics = ['accuracy']
```

Fitted Model Results



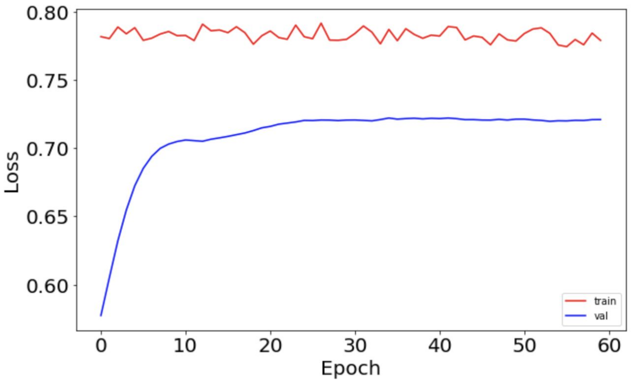
```
Accuracy on training data: 0.5503221154212952%
 Error on training data: 0.44967788457870483
Accuracy on test data: 0.5479224324226379%
 Error on test data: 0.45207756757736206
```

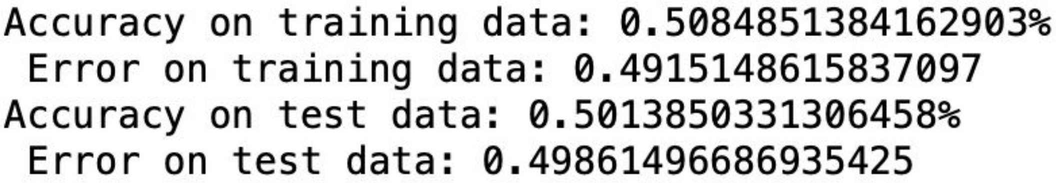


```
Accuracy on training data: 0.24416430294513702%
 Error on training data: 0.755835697054863
Accuracy on test data: 0.23490305244922638%
 Error on test data: 0.7650969475507736
```

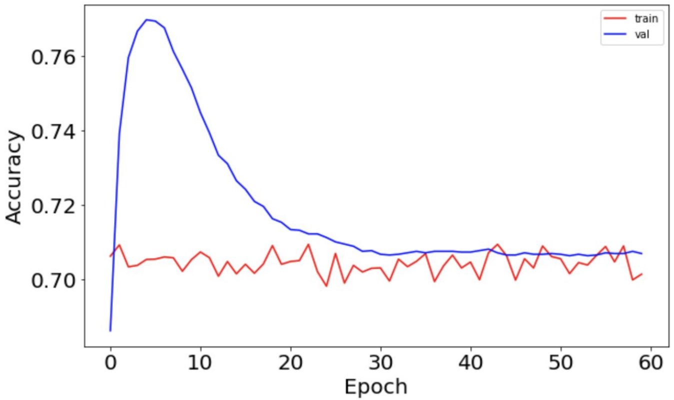
Stochastic Gradient Descent: Genetic Disorder

The learning rate decreases according to this function: lr=lr×1/(1+decay∗epoch)



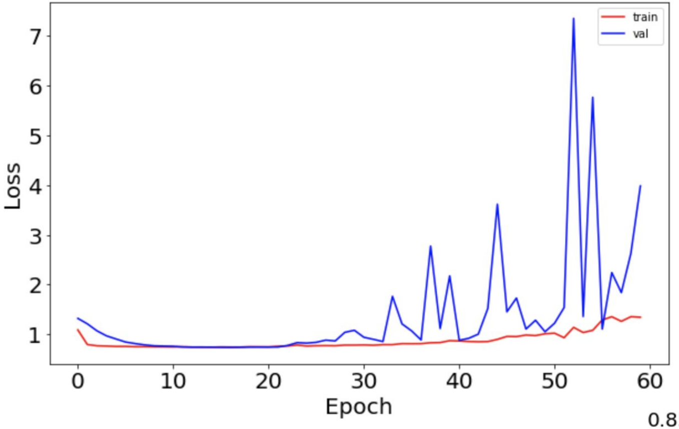


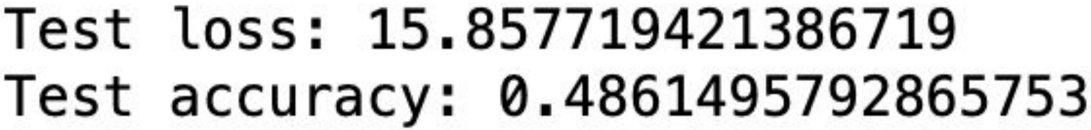
```
Accuracy on training data: 0.5084851384162903%
 Error on training data: 0.4915148615837097
Accuracy on test data: 0.5013850331306458%
 Error on test data: 0.49861496686935425
```



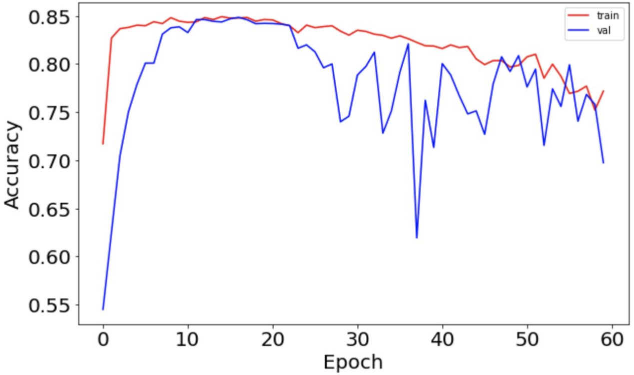
Stochastic Gradient Descent: Disorder Subclass

The learning rate decreases according to this function: lr=lr×1/(1+decay∗epoch)



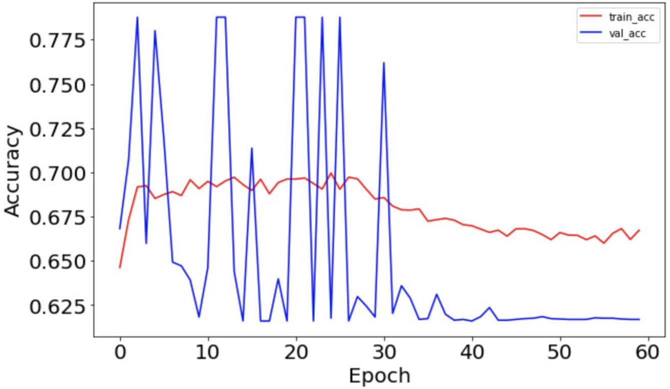


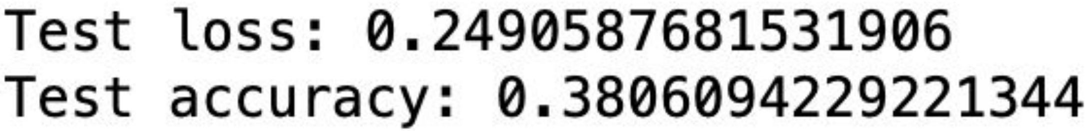
```
Test loss: 15.857719421386719
Test accuracy: 0.4861495792865753
```



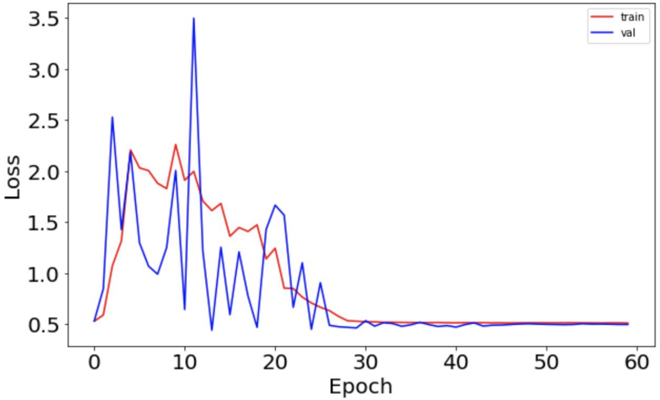
Exponential Decay: Genetic Disorder

exponential decay model to experiment with a different learning rate function: $lr=lr_0 \times e$^$(-kt)$



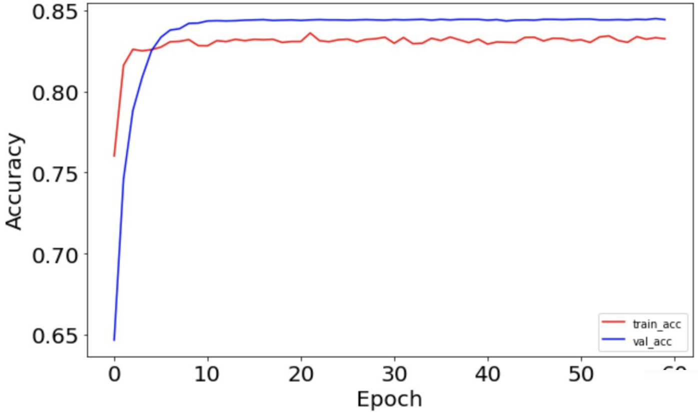


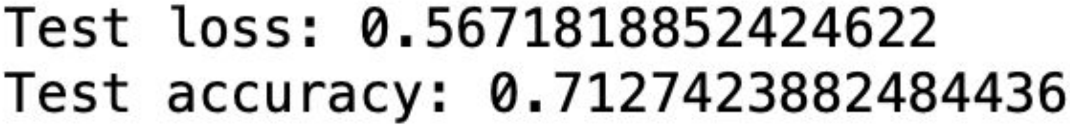
```
Test loss: 0.2490587681531906
Test accuracy: 0.3806094229221344
```



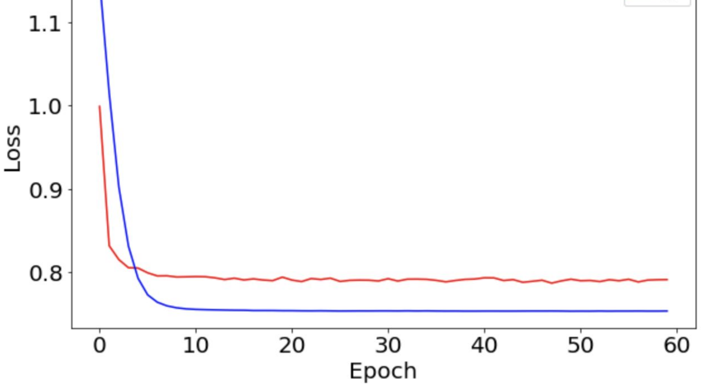
Exponential Decay: Disorder Subclass

exponential decay model to experiment with a different learning rate function: lr=lr_0 × e^(−kt)





```
Test loss: 0.5671818852424622
Test accuracy: 0.7127423882484436
```



Conclusions & Improvements

- The best optimizer for the Genetic Disorder Model is the Adam optimizer
- The best optimizer for the Disorder Subclass is the SGD with the Exponential Decay Learning Rate Scheduler

- Neither model achieved very substantial accuracy but the Disorder Subclass model was more successful than the Genetic Disorder
- More hard genetic data, and more data across the whole data set will increase predictive power