HUNTINGTONS DISEASE PREDICTION

Varshaa Shree Bhuvanendar (G01269710), Deeksha Gangadharan Srinivas (G01291097)

December 12, 2021

Abstract

We propose a classification approach to determine Huntington's Disease in a patient given data such as Age, Sex and protein values. We applied Decision Tree, Random Forest, and Generalized Linear Model on the given data. We found 41 genes that contribute towards Huntington's Disease. Generalized Linear model performs best on the data followed by Random Forest and Decision Tree with accuracies of 99.4,96.8 and 92.6 respectively. Cross-validation is applied for Random Forest and Decision Tree. The reason for choosing the above algorithms being decision trees and random forests can handle large datasets, also these algorithm can handle both categorical and numerical data. GLM(Generalized Linear Model) generalized linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. .Apart from this we also analyzed dataset with limited number of rows with gene sequences of 12 patients and applied naive bayes algorithm to find out that cytosine, one of the four nucleobases forming the DNA, contributes highest for the cause of Huntington's disease. Through the project we aimed to explore the impact of Machine Learning on Huntington's Patient data which could be of two types first it could consist of protein valued data or it could contain gene sequences. Most of the research currently performed is done on gene sequenced data. So, we explored protein-based data and how different machine learning models adapt to this data.

1. INTRODUCTION

Huntington's disease is a rare, inherited disease that causes the progressive breakdown (degeneration) of nerve cells in the brain. It has a broad impact on a person's functional abilities and usually results in movement, thinking (cognitive), and psychiatric disorders. We are analyzing the genomic data to study the features that cause Huntington's Disease from data obtained from NCBI (National Center for Bioinformatics) using machine learning techniques such as Decision Tree, Random Forest, and Generalized Linear Model.

The time from disease emergence to death is often about 10 to 30 years. Juvenile Huntington's disease usually results in death within 10 years after symptoms develop. **Therefore, it becomes important to detect this disease as early as possible so that we can prevent it from progressing at a higher rate with proper medication.**

How is our project different from other papers?

- In other papers they have used gene sequence we have used protein data. We have also shown activation value for each contributing protein.
- Also, most papers that exist have explored clustering techniques. We have tried classification techniques.

2. PROBLEM STATEMENT

PART 1:

We aim to analyze data obtained from NCBI (National Center for Biotechnology Information) which consists of information related to people with and without Huntington's disease the features included Age,

Sex, and different protein values like SERPINA3, ACTN2, etc. Using the above dataset, we determined the most prominent proteins that contributed to Huntington's Disease determine the most prominent contributing genes towards Huntington's disease.

Each gene can contribute towards a disease, we aim to determine which genes contribute towards Huntington's disease and what is the value that triggers Huntington's disease.

PART 2:

Part two is an exploration. We found a dataset with limited set of rows of gene sequences of 12 Huntington's Disease patient we tried to find which chemical base in a gene contributes the most towards Huntington's Disease.

2.1. Notations

- GLM: Generalized Linear Model
- NCBI: National Center for Bioinformatics

3. LITERATURE REVIEW

During our research on Huntington's disease we came across papers where they used clustering to find commonality among Huntington's disease patient. In April 2013 Jinnie Ko, Hannah Furby, Xiaoye Ma, Jeffrey D. Long, Xiao-Yu Lu, Diana Slowiejko, Rita Gandhy published a paper in which the used KMeans to cluster the data considered included demographic, genetic/medical/family history, social, symptom and medication-use values.

Another paper that we referred was Machine learning spots the time to treat Huntington disease. It was published in February 2021 by Thomas Dighiero-Brecht, Allan Tobin, and Sarah Tabrizi. In this paper they used HD Course Map on the Data.

4. METHODS AND TECHNIQUES

To determine the possible genes that contribute towards Huntington's disease we use Decision Tree, Random Forest, and Generalized Linear Model.

1) Decision Tree:

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

In our case, the decision tree would provide various genes that contribute towards Huntington's Disease.

For decision tree we considered criterion ='entropy',max_depth = 4,min_samples_split = 4 provided the best results.

How do we decide which genes contribute? At every step of the decision tree, we choose a node, this node that we choose for split as the tree progress at leaf there can be only two possibilities having Huntington's Disease Yes or No. For nodes that give Yes, we see which have the highest number of items or rows. We choose all nodes leading to this path.

Next to decide activation values based on values mentioned within the nodes for each gene. These values act as the activation values for each gene.

2) Random Forest:

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

For random forest we took n_estimators=90,max_depth = 4,criterion='entropy' gave out the best accuracy.

How do we find the most prominent genes in Random Forest?

We choose the best fit tree next we perform the same procedure as the decision tree. We choose the leaf node with the most rows classified the backtrack to root.

Cross-validation has been applied to both decision trees and Random Forest algorithm.

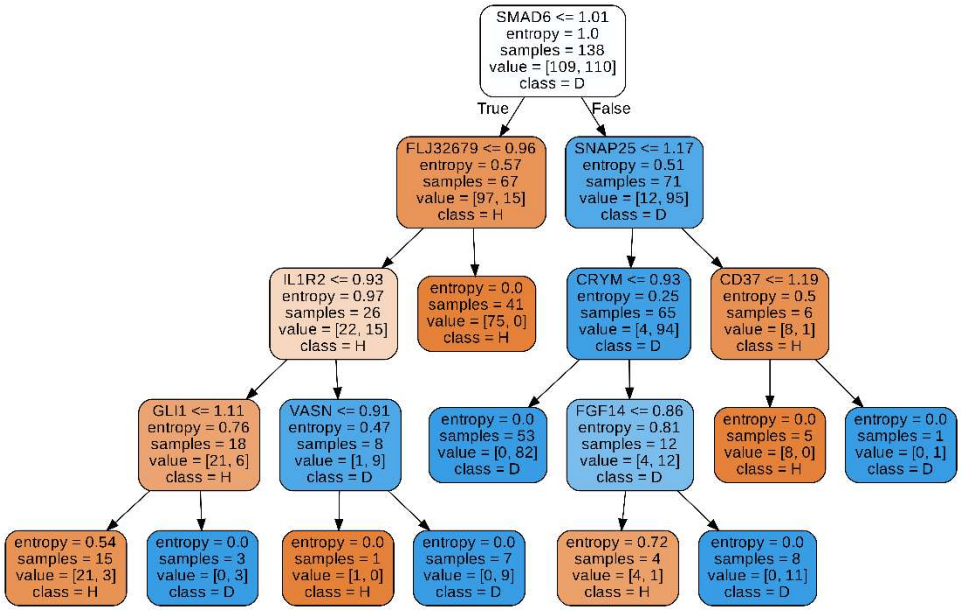


Fig 1: The above tree depicts each protein chosen as node for the creation of tree at different layers and value of split for each branch

3) Generalized Linear Model:

GLMs can be used to construct the models for regression and classification problems by using the type of distribution which best describes the data or labels given for training the model.

To construct GLMs for a particular type of data or more generally for linear or logistic classification problems the following three assumptions or design choices are to be considered:

$$y|x;\theta \sim exponential\ family(\eta)$$
$$Given\ x\ our\ goal\ is\ to\ predict\ T(y)\ which\ is\ equal\ to\ y\ in\ our\ case\ or\ h(x) =$$
$$E[y|x] = \mu$$
$$\eta = \theta^T * x$$

GLM determines predictors (the gene profiling) with non-zero coefficients indicating a linear contribution of the gene profiling to the prediction.

A cross-validation strategy was used to train the GLM and to evaluate its performance

GLM Identified 14 genes with non-zero coefficients

Reason for choosing above algorithms:

Decision Trees:

- Easy to use and understand.
- Can handle both categorical and numerical data.
- Resistant to outliers, hence requiring little data preprocessing.

Random Forest:

- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion and our dataset has a large amount of columns.
- It gives estimates of what variables are important in the classification.
- It has an effective method for estimating missing data and maintains accuracy when a substantial proportion of the data are missing.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- It offers an experimental method for detecting variable interactions.

Generalized Linear Model:

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

PART 2:

In the second part we imported data of the 12 Huntington's Disease patients. We applied Naïve Bayes algorithm on data and determined which nucleobase contributes the highest or occurs the most in the sequence to contribute towards Huntington's Disease

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

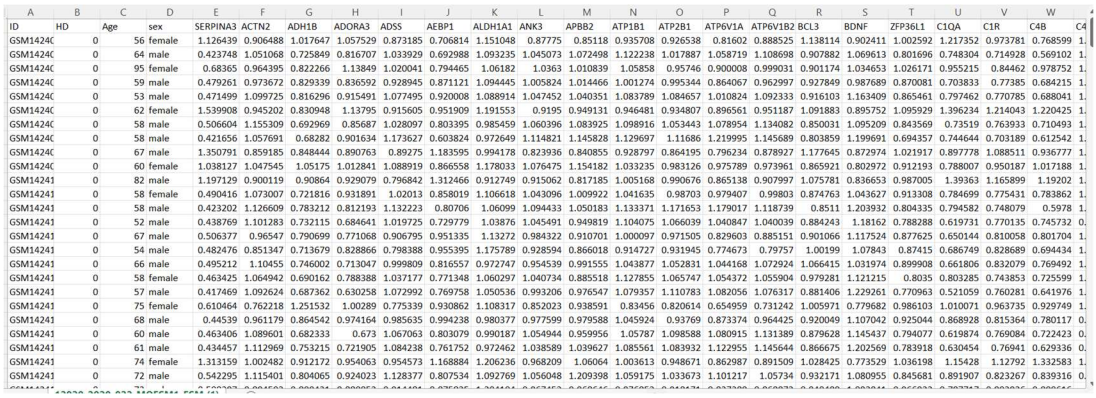
5. DISCUSSION AND RESULTS

5.1 Datasets:

Both the datasets were obtained from NCBI (National Center for Biotechnology Information)

PART 1: Two datasets are being used for the prediction of Huntington's disease and to identify the contributing proteins with their respective values for which this disease is triggered:

Dataset: This Dataset Contains various protein values (SERPINA3, ACTN2, etc.) along with common parameters like age and gender. This dataset is used to determine/identify those proteins that are contributing to Huntington's disease.



ID	HD	Age	sex	SERPINA3	ACTN2	ADH1B	ADORA3	ADSS	AEBP1	ALDH1A1	ANK3	APBB2	ATP1B1	ATP2B1	ATP6V1A	ATP6V1B2	BCL3	BDNF	ZFP36L1	C1QA	C1R	C4B	C4
GSM1424(	0	56	female	1.126439	0.906488	1.017647	1.057529	0.873185	0.706814	1.151048	0.87775	0.85118	0.935708	0.926538	0.81602	0.888525	1.138114	0.902411	1.002592	1.217352	0.973781	0.768599	1.
GSM1424(	0	64	male	0.423748	1.051068	0.725849	0.816707	1.033929	0.692988	1.093235	1.045073	1.072498	1.122238	1.017887	1.058719	1.108698	0.907882	1.069613	0.801696	0.748304	0.714928	0.569102	1.
GSM1424(	0	95	female	0.68365	0.964395	0.822266	1.13849	1.020041	0.794465	1.06182	1.0363	1.010839	1.05858	0.95746	0.900008	0.999031	0.901174	1.034653	1.026171	0.955215	0.84462	0.978752	1.
GSM1424(	0	59	male	0.479261	0.973672	0.829339	0.836592	0.928945	0.871121	1.094445	1.005824	1.014466	1.001274	0.995344	0.864067	0.962997	0.927849	0.987689	0.870081	0.703833	0.77385	0.684215	1.
GSM1424(	0	53	male	0.471499	1.099725	0.816296	0.915491	1.077495	0.920008	1.088914	1.047452	1.040351	1.083789	1.084657	1.010824	1.092333	0.916103	1.163409	0.865461	0.797462	0.770785	0.688041	1.
GSM1424(	0	62	female	1.539908	0.945202	0.830948	1.13795	0.915605	0.951909	1.191553	0.9195	0.949131	0.946481	0.934807	0.896561	0.951187	1.091883	0.895752	1.095929	1.396234	1.214043	1.220425	1.
GSM1424(	0	58	male	0.506604	1.155309	0.692969	0.85687	1.028097	0.803395	0.985459	1.060396	1.083925	1.098916	1.053443	1.078954	1.134082	0.850031	1.095209	0.843569	0.73519	0.763933	0.710493	1.
GSM1424(	0	58	male	0.421656	1.057691	0.68282	0.901634	1.173627	0.603824	0.972649	1.114821	1.145828	1.129697	1.11686	1.219995	1.145689	0.803859	1.199691	0.694357	0.744644	0.703189	0.612542	1.
GSM1424(	0	67	male	1.350791	0.859185	0.848444	0.890763	0.89275	1.183595	0.994178	0.823936	0.840855	0.928797	0.864195	0.796234	0.878927	1.177645	0.872974	1.021917	0.897778	1.088511	0.936777	1.
GSM1424(	0	60	female	1.038127	1.047545	1.05175	1.012841	1.088919	0.866558	1.178033	1.076475	1.154182	1.033235	0.983126	0.975789	0.973961	0.865921	0.802972	0.912193	0.788007	0.950187	1.017188	1.
GSM14241	0	82	male	1.197129	0.900119	0.90864	0.929079	0.796842	1.312466	0.912749	0.915062	0.817185	1.005168	0.990676	0.865138	0.907997	1.075781	0.836653	0.987005	1.39363	1.165899	1.19202	1.
GSM14241	0	58	female	0.490416	1.073007	0.721816	0.931891	1.02013	0.858019	1.106618	1.043096	1.009922	1.041635	0.98703	0.979407	0.99803	0.874763	1.043627	0.913308	0.784699	0.775431	0.783862	1.
GSM14241	0	58	male	0.423202	1.126609	0.783212	0.812193	1.132223	0.80706	1.06099	1.094433	1.050183	1.133371	1.171653	1.179017	1.118739	0.8511	1.203932	0.804335	0.794582	0.748079	0.5978	1.
GSM14241	0	52	male	0.438769	1.101283	0.732115	0.684641	1.019725	0.729779	1.03876	1.045491	0.949819	1.104075	1.066039	1.040847	1.040039	0.884243	1.18162	0.788288	0.619731	0.770135	0.745732	0.
GSM14241	0	67	male	0.506377	0.96547	0.790699	0.771068	0.906795	0.951335	1.13272	0.984322	0.910701	1.000097	0.971505	0.829603	0.885151	0.901066	1.117524	0.877625	0.650144	0.810058	0.801704	1.
GSM14241	0	54	male	0.482476	0.851347	0.713679	0.828866	0.798388	0.955395	1.175789	0.928594	0.866018	0.914727	0.931945	0.774673	0.79757	1.00199	1.07843	0.87415	0.686749	0.828689	0.694434	1.
GSM14241	0	66	male	0.495212	1.10455	0.746002	0.713047	0.999809	0.816557	0.972747	0.954539	0.991555	1.043877	1.052831	1.044168	1.072924	1.066415	1.031974	0.899908	0.661806	0.832079	0.769492	1.
GSM14241	0	58	female	0.463425	1.064942	0.690162	0.788388	1.037177	0.771348	1.060297	1.040734	0.885518	1.127855	1.065747	1.054372	1.055904	0.979281	1.121215	0.8035	0.803285	0.743853	0.725599	1.
GSM14241	0	57	male	0.417469	1.092624	0.687362	0.630258	1.072992	0.769758	1.050536	0.993206	0.976547	1.079357	1.110783	1.082056	1.076317	0.881406	1.229261	0.770963	0.521059	0.760281	0.641976	1.
GSM14241	0	75	female	0.610464	0.762218	1.251532	1.00289	0.775339	0.930862	1.108317	0.852023	0.938591	0.83456	0.820614	0.654959	0.731242	1.005971	0.779682	0.986103	1.010071	0.963735	0.929749	1.
GSM14241	0	68	male	0.44539	0.961179	0.864542	0.974164	0.985635	0.994238	0.980377	0.977599	0.979588	1.045924	0.93769	0.873374	0.964425	0.920049	1.107042	0.925044	0.868928	0.815364	0.780117	0.
GSM14241	0	60	male	0.463406	1.089601	0.682333	0.673	1.067063	0.803079	0.990187	1.054944	0.959956	1.05787	1.098588	1.080915	1.131389	0.879628	1.145437	0.794077	0.619874	0.769084	0.722423	0.
GSM14241	0	61	male	0.434457	1.112969	0.753215	0.721905	1.084238	0.761752	0.972462	1.038589	1.039627	1.085561	1.083932	1.122955	1.145644	0.866675	1.202569	0.783918	0.630454	0.76941	0.629336	0.
GSM14241	0	74	female	1.313159	1.002482	0.912172	0.954063	0.954573	1.168884	1.206236	0.968209	1.06064	1.003613	0.948671	0.862987	0.891509	1.028425	0.773529	1.036198	1.15428	1.12792	1.332583	1.
GSM14241	0	72	male	0.542295	1.115401	0.804065	0.924023	1.128377	0.807534	1.092769	1.056048	1.209398	1.059175	1.033673	1.101217	1.05734	0.932171	1.080955	0.845681	0.891907	0.823267	0.839316	0.

Fig 2: Dataset with each protein and contribution value

PART 2: We obtained some minimal about 12 patient's data of CAG sequence which are the protein in GENE of Huntington patients

	rs1065745*		rs34315806*		rs363099*		rs362336*		rs362331		rs362273*		rs149109767		rs2276881		rs362272		r362307		Haplotype	
Sample	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L
GM01169f	C	C	C	C	T	C	A	G	C	T	G	A	D	D	G	G	A	G	C	T	8	1
GM02077f	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	T	3	1
GM02079f	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	T	2	1
GM02147f	C	C	C	C	T	C	A	G	C	T	G	A	D	D	G	G	A	G	C	C	8	6
GM02151f	C	C	C	C	T	C	A	G	C	T	G	A	D	D	G	G	A	G	C	C	8	6
GM03866f	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	T	2	1
GM21756f	C	C	C	C	T	C	A	G	C	T	G	A	D	D	G	G	A	G	C	C	8	3
ND30259	C	C	T	C	C	C	G	G	T	C	A	G	D	D	G	G	G	G	C	C	other	6
ND30626	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	C	11	2
ND31038	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	C	2	2
ND33947	C	C	C	C	T	C	A	G	C	T	G	A	D	D	G	G	A	G	C	T	8	1
GM13505f*	C	C	C	C	C	C	G	G	T	T	A	A	D	D	G	G	G	G	C	C	NA	NA

Fig 3: Gene sequence data

5.2 Evaluation Metrics

Accuracy was used as the measure to conclude on the best Machine Learning Algorithm for predicting Huntington's Disease.

Decision Tree yielded an accuracy of 0.92 and with Random Forest we obtained an accuracy of 0.96 and GLM (Generalized Linear Model) gave the highest accuracy of 0.99.

```
[ ] import matplotlib.pyplot as plt
    x = ["Decision Tree","Random Forest","General Linear Models"]
    y =[str(decision_tree_accuracy),str(random_forest_accuracy),str(glm_accuracy['train'])]
    plt.plot(x, y)
    plt.show()
```

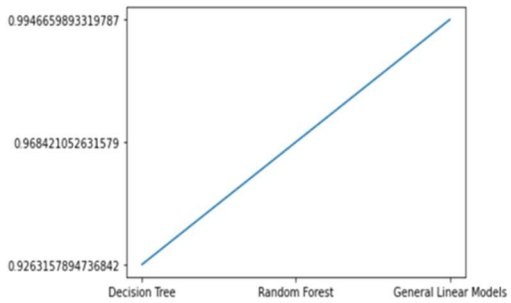


Fig 4: Accuracy plot

5.3 Experimental Results

```
[124] r= fit_glm.coef_norm()

      d = dict((k, v) for k, v in r.items() if v > 0)
      d
      pd.Series(d.values(), d.keys()).plot(kind='barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fce2daa7bd0>
```

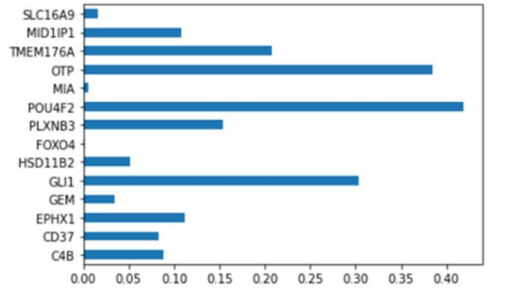


Fig 5: Gene contribution from GLM: POU4F2 has the highest contribution of 0.42

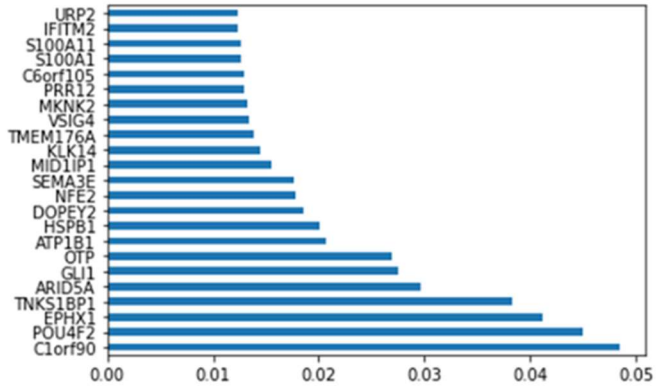


Fig 6: Genes Identified using Random Forest: POU4F2 has the highest contribution of 0.48

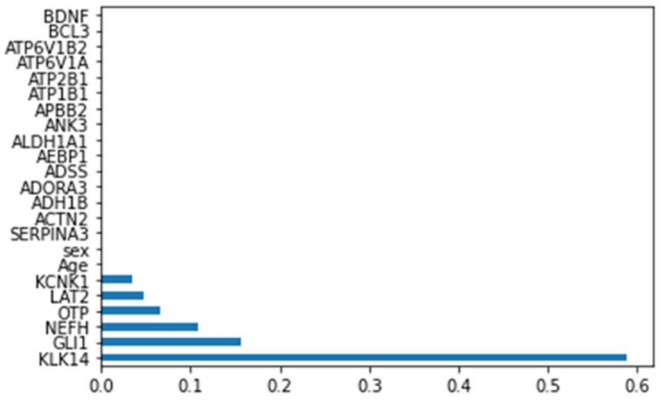


Fig 7: Genes Identified using Decision Tree: KLK14 has the highest contribution

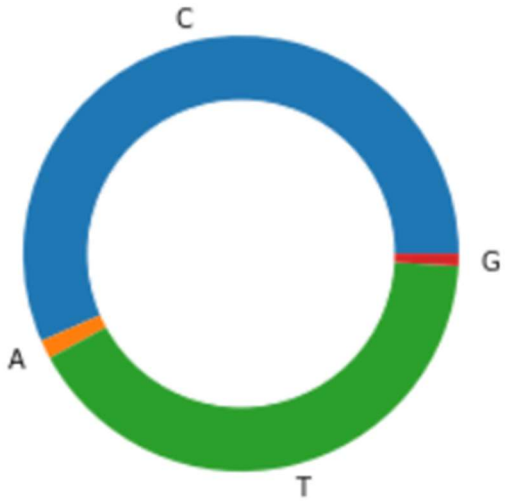


Fig 8: As we can see in the donut chart C: Cysteine has the largest contribution among all chemical bases within the gene

6. CONCLUSION

We conclude that there are 41 genes in total that we found common among all algorithms that are contributing to Huntington's Disease.

Among the three algorithms, we used GLM is the best for the prediction of Huntington's Disease with an accuracy of close to 99% for the dataset provided. Second, to GLM Random Forest performs better with an accuracy of 96%.

Each of these algorithms gives out a set gene which they feel best contribute towards Huntington's disease

6.1 Directions for Future Work

Machine Learning Algorithms have been used in very little research in this area. We employed a Decision Tree, Random Forest Algorithm, and General Linear Models in this project. In the future, we can investigate more algorithms to see which one is most suitable and produces the best results for Huntington's disease prediction. Furthermore, because there was a lack of data, it is critical to document data that is more specific to Huntington's Disease. Furthermore, this dataset included more columns and fewer rows, indicating that it contained a significant number of attributes and fewer patients. As a result, the model lengthened the training period. This dataset comprises a variety of protein values as well as other information such as age and gender. We're using this data to figure out which proteins have a role in Huntington's disease. Our model, on the other hand, does not predict or calculate the activation value of the protein involved in Huntington's Disease. As a result, understanding the role of each protein in Huntington's Disease is critical.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2877728/

[2] https://www.kaggle.com/zakarii/dna-sequence-classification-cnn-gru

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534

[4] https://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-7-8

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC134256/

[6] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000222.v3.p2