IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Machine Learning-Assisted Prediction of Variant Pathogenicity in Neuro-Genetic Disorders

Ishita Rao[1], Nirupma Singh[2]

1. The International School, Bangalore, Karnataka, India; ishitarao08@gmail.com
2. Department of Biological Sciences and Engineering, Netaji Subhas University of Technology (Formerly NSIT), Dwarka, Delhi, India

ABSTRACT: Neurological disorders and their underlying genetic causes continue to be vastly unexplored. However, with the advent of technology, there is an increased scope for analyzing at the genetic level by looking at variants, cell functions, and risk alleles. In this study, from the GWAS (Genome Wide Association Study) catalogue and existing data sets from studies, single-nucleotide variants (SNVs) in terms of risk allele, mapped genes, traits, and p-value associated with specific neurological disorders were collected, followed by gene and variant identification. Using network and pathways analysis software STRING, Cytoscape, and the Reactome database, top genes were identified. The risk alleles associated with the genes of SNVs were taken and input into Ensembl's Biomart and annotated in Ensembl's VEP (Variant Effect Predictor). This collated information helped in gaining more information about the associated SNVs and helped identify the impact of these variants on the disorders. The ML models, RidgeCV and HuberRegressor, trained on the variant annotated dataset to make a risk score prediction system and classification system, achieved highly accurate results. This prediction system can significantly aid in clinical and scientific analysis, which, in the future, may result in the betterment of patient care and treatment.

KEYWORDS: Neurological Disorders, Machine Learning, Structural Variants, Single Nucleotide Variants, and Risk Scoring.

Introduction

Neurological disorders, in summary, are all conditions related to the central and peripheral nervous system. These disorders can be categorized in countless ways. A few categories include neurodegenerative, neurodevelopmental, neuromuscular, neuropsychiatric, seizure, cerebrovascular, and infectious disorders. According to the World Health Organization (WHO), neurological disorders are the second leading cause of death[1] with approximately 9 million deaths annually.[2] In India specifically, there is a huge prevalence of neurological disorders as per the Indian Council for Medical Research (ICMR). An example is India's contribution of 20 percent of the global cases of epilepsy.[3] There are a lot of external and internal factors that influence these disorders, but their causes are still a vastly uncharted domain.

In the small sample size of research for this project, it was discovered that most studies focused on uncovering the genetic causes of specific symptomatic neurological disorders. For instance, in a study of muscular system atrophy, they identified novel risk loci,[4] or in a study of Alzheimer's, where they tried to understand the SNPs and their impact on proteins.[5] Additionally, in a study about Amyotrophic lateral sclerosis(ALS), they uncovered changes in terms of structural variations across 25 known ALS genes,[6] and in research about Multiple sclerosis, they identified 10 candidate nsSNPs and assessed their functional impact.[7] In a study about Dyslexia,[8,9] Childhood Apraxia of Speech (CAS),[10] and Intellectual Disabilities,[11] they performed wide genome analysis and identified [12] pathogenic or likely pathogenic variants in key developmental genes. The diversity and complexities of such conditions allow for a lot of scope for research in this area of mental health illnesses.

Research into the genetic architecture of neurological disorders is expanding. Studies have found multiple genetic variations and mutations, such as structural variants (SVs), single-nucleotide variations (SNVs), short tandem repeat (STR) variants,[12] and copy-number variants in the exome that could suggest pathogenic properties with respect to these disorders. With the rise of multiple bioinformatics tools and Genome Sequencing technology like Oxford Nanopore PromethION platform,[13] SV/CNV calling (Sniffles, CuteSV), Pathogenicity prediction through PolyPhen-2, SNPs & Go, PredictSNP, variant sourcing from the NHGRI-EBI GWAS Catalog, Ensembl VEP to characterize variant effects, and so many more, our understanding has broadened.

With reference to the research collected, it was understood that there was no clear method that could be used to identify risk alleles and variants, and which disorder they could lead to. Understanding if there are monogenic risk loci that are associated with such conditions can help in the early diagnosis and treatment of patients in this age range, also preventing any further neurological deterioration or reduced development. Therefore, the hypothesis for this study is to analyze existing databases and determine if there are monogenic causes in the form of structural variants and single-nucleotide variants for neurological disorders, and create a detection system that can show from a given dataset whether the risk allele or variant provided can be benign or malignant.

DOI: 10.36838/IJHSR813.6

Methods

Data collection and processing:

For the data collection part of the study, the GWAS (Genome Wide Association Study) Catalogue was explored. Keywords related to neurological and neuropsychological diseases, such as Parkinson's disease, Bipolar disorder, Major Depressive Disorder (MDD), obsessive-compulsive disorder, Borderline Personality Disorder, Amyotrophic Lateral Sclerosis (ALS), and Mental health disorders were used in the search. A list of the search results and information about the studies was compiled. From this record, based on the number of gene associations and traits, we shortlisted specific studies. From these TSV files of the data from the study, the columns that displayed the mapped genes for the disorder, the P-value, possible risk-alleles, and names of the traits associated were compiled separately in a single data file. This led to us attaining a final gene list. During this project, various online processing tools were utilized to get this information, such as STRING version 12.0, Reactome version 94, gnomAD version 4.1, and Ensembl VEP 115. In addition to this, the project accessed datasets from Clinvar in October 2025.

Network analysis:

The final gene list was input and processed into the STRING database. The interactions for all the genes were identified, and a gene-interaction network was given by the STRING, as shown in Figure 1. The network made use of clear lines, nodes, and colors to show the relation each gene had with the others. It was understood from the information that the connections between the genes were in regard to their involvement in functions and processes in neurological diseases. The data from the network was exported in a file and then opened in the network analysis application software Cytoscape. Using a tool in Cytoscape called 'Analyse Network', a diagram with all the relevant nodes and tables with the necessary topological parameters was computed. The node table provided data on how many nodes each gene was connected with and quantified the influence of each gene with its respective node in terms of centrality. After this, the parameters degree, betweenness centrality, closeness centrality, clustering coefficient, and name were selected and placed into a separate file. This was to make the data decluttered, as it allowed us to see the properties of the genes in the network clearly. Based on the highest degree of each node, the data in the file was sorted. The top genes based on this parameter were selected and kept aside. This was important because we could then further explore and examine the genes that were most strongly interconnected in the network of genes related to neurological diseases.

Pathway analysis:

The Reactome database was then effectively used to identify specifically which biological pathways and processes are most involved with respect to the genes from the gene list compiled after network analysis. Reactome produced a table with all the pathway names and biological functions listed. The application also provided results for genes that were found to have no involvement in pathways, as well as a thorough report of its analysis. This was crucial in documenting and identifying all the necessary genes impacting cellular and biological processes. The top genes identified previously were then searched for in the table to shortlist further. If a particular gene was present among the top 25 pathways, it was singled out and highlighted.

Variant annotation:

The risk alleles associated with the final gene list were taken and annotated. The rs IDs of each gene were input into Ensembl's Biomart feature for human single-nucleotide variants. From these, specific parameters about associated SNV variants with the risk alleles, different scores, gene, protein, chromosome, allele, and clinical significance, such as SIFT score and prediction, Gene name, P-value, Chromosome scaffold position start and end, Variant alleles, and any associated SNV variant risk allele name were selected. This led to the extraction of vital genetic information about the risk alleles. Following this, the risk alleles were input into the Variant Effect Predictor (VEP) tool of Ensembl to understand the neurological and genetic effects of the identified variants on neurological diseases. A collated dataset with specific parameters from the results of VEP and Biomart was then formed for further analysis. Additionally, ClinVar and gnomAD databases were used to carry out the enrichment and annotation for the structural variants identified for neurological disorders from the dbVar database. The parameters, such as chromosome location, start and end position, the type of SVs, gene associated, allele frequency, ClinVar condition, and ClinVar significance, were identified.

Model building and evaluation:

Two algorithms were written to create a genetic classification model, one for the Single Nucleotide Variant (SNV) annotated variant data set and one for the Structural Variant (SV) annotated variant data set. The algorithms created a regression model, which is a supervised machine learning model that can be used to predict a risk score and value for an input of neurogenetic and variant data. The two models first load all the libraries and datasets. The algorithms read their respective datasets and then convert the column Clinical risk score to a numeric risk score using a 0-1 scale: 0.00 is assigned if the gene is benign, 0.25 if it is likely benign, 0.50 if there is uncertain significance of its effect, 0.75 if likely pathogenic, and 1.00 if it is pathogenic. The algorithm then goes on to define the features of the data set, and the code should output where it is divided into numerical and categorical features. For the creation of the SNV regression model, the project used the categorical parameters Consequence which classifies the type of effect on the gene structure, Impact which suggests the severity of the consequence, Biotype which details the functional category of the gene, SIFT prediction which predicts if a missense mutation is harmful, and PolyPhen prediction which produces a risk score for the mutation's degree of damage. The included numerical parameters were CADD PHRED, which produces a quality score on the variants, CADD Raw, which shows the original output of the model, REVEL, which is cru-

cial in predicting pathogenicity of the variants, and Clinical Prediction, which uses the ACMG classification to categorize the variants. Similarly, the SV model included the categorical parameters Structural Variant Type, Gene gnomAD, which shows its associated gene, and ClinVar database Condition. The included numerical parameters were Start, End, and AF values of the variant. These parameters help in the effective classification of the variants and in creating an accurate variant scoring system. The models then reduce high cardinality categorical features, where they only keep the top 30 categories and label the rest as "others." This is done to make the model more efficient and generalizable because one-hot encoding, which converts each category to binary columns, would otherwise create numerous new columns.

Both models then define their data preprocessing pipelines. For the numerical pipeline, the code performs 2 steps: median imputation, where a missing value in a dataset is replaced by the median of that column, and standard scaling, which rescales all the numeric columns to have the median as 0 and standard deviation as 1, thus making it easier to train the model. In the same way, there are 2 steps for the categorical features - constant imputation, which fills missing values as "unknown," and one-hot encoding.

The 2 codes use two linear regression models: HuberRegressor, a model robust to outliers, and RidgeCV, a ridge regression with internal cross-validation. It trains and evaluates these models for the respective SNV and SV datasets by running a 5-fold cross-validation using the R^2 score, a measure that indicates the variation in the dependent variable. The 5- fold cross-validation was stratified, ensuring the same percentage of samples was provided. This maintained the validity and accuracy of the model during training. The code calculates from the model and reports the mean, standard deviation, test R^2 value, test mean absolute error, and suggests the best model with the highest mean CV R^2 (Cross-validation R^2). It saves the information and provides a visualization of the R^2 scores as a boxplot, and shows the true vs the predicted risk scores for the best model. The algorithms for SNV and SV extract the features from the preprocessing pipeline, retrieve coefficients for the models, and plot the top 20 most important features by the absolute coefficient value. This is important because it helps identify what features (categorical and numerical) have the biggest impact on the risk score produced for the genetic variants. Through this, an understanding of what the model learned is gained, and the model also provides insight into which biological features are predictors of pathogenicity, which is crucial and has a lot of clinical relevance.

■ Results and Discussion

Network and pathway analysis:

After data processing, a list of 998 genes was retained. On placing this list into the STRING database, the interactions for the genes were identified, and a network was made as shown in Figure 1. The STRING database also produced a gene ontology bubble plot, as seen in Figure 2, which showed the most represented neurological processes in the gene set network. Following this, the network was visualized in Cytoscape, and it had 524 nodes and 1773 edges. After network analysis, from the computed topological parameters the following degrees were selected - degree which measured the frequency of the gene in the network, gene name, Betweenness Centrality which measures the node's reliance by quantifying how often it lies on the shortest path between other nodes, Closeness Centrality which indicates a node's efficiency by measuring its average shortest distance to all other nodes in the network and Clustering Coefficient which measures how interconnected the neighbors of a node are in a network. Thus, on the basis of these parameters, the top genes were identified. The top 10 genes out of all are listed in Table 1. These are genes that are neurologically significant to disorders such as Alzheimer's disease, Parkinson's disease, and the neurodevelopmental disorder intellectual disability.

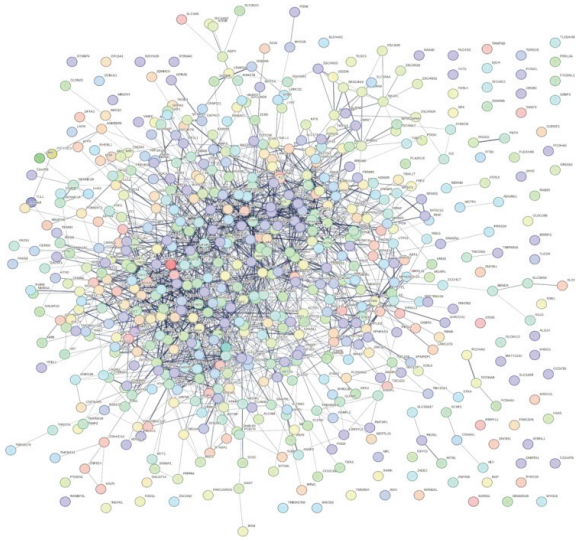

Figure 1: Image of a neurological gene network. The network was created based on a medium confidence degree of 0.400. The Network nodes represent proteins, splice isoforms, or post-translational modifications that are collapsed, i.e., each node represents all the proteins produced by a single protein-coding gene locus. 524 node colors represent individual query proteins, where the 3D structure is known or predicted, and the first shell of interactors. The white nodes are separate and represent the second shell of interactors. In addition to this, there are empty nodes with proteins of unknown 3D structure.

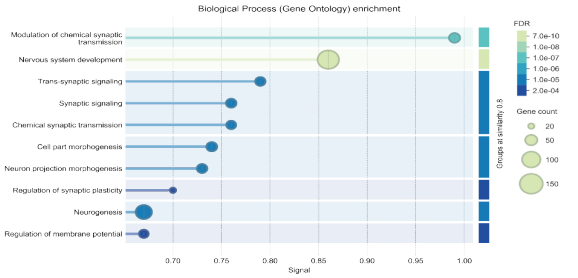

Figure 2: Image of gene-related biological process enrichment. The figure here is a bubble plot, which shows the different biological processes against the Y-axis (GO terms), where each horizontal bar corresponds to one GO biological process category. In the X-axis, the enrichment score (often referred to as the "rich factor"), which is the ratio of the number of genes in your set associated with a GO term vs. the total number of genes associated with that term in the background/reference genome. Here, the bubble size represents the gene count, a value that shows how many of the genes are associated with that process; thus, a larger circle indicates more genes. Additionally, the bubble color represents the FDR (False Discovery Rate) significance level, where the lighter green/yellow are highly significant (very low FDR), and darker blue are less significant (higher FDR, but still below threshold).

DOI: 10.36838/IJHSR813.6

Table 1: The top 10 genes related to neurological disorders.

Serial Number	Name	Degree	Betweenness Centrality	Closeness Centrality	Clustering Coefficient
1.	*EP300*	60	0.160591805	0.3971742543	0.09548022599
2.	*ESR1*	45	0.08218605988	0.3815987934	0.1151515152
3.	*CACNA1C*	44	0.04740246773	0.3677325581	0.1501057082
4.	*ANK3*	44	0.08011472192	0.3750926612	0.09725158562
5.	*MAPT*	37	0.06722154708	0.386259542	0.1456456456
6.	*GRIA1*	35	0.03706845225	0.359886202	0.1831932773
7.	*SNCA*	35	0.04831354062	0.359886202	0.1529411765
8.	*GRIN2A*	32	0.02077170651	0.3513888889	0.2217741935
9.	*LRRK2*	32	0.05118771169	0.3583569405	0.1169354839
10.	*PRKCA*	29	0.02955994534	0.3642908567	0.1502463054

Following the pathway analysis in Reactome, 1343 pathway names and functions were identified. The top 5 pathways identified were Meiosis (R-HSA-1500620), Meiotic synapsis (R-HSA-1221632), B-WICH complex positively regulates rRNA expression (R-HSA-5250924), Defective pyroptosis (R-HSA-9710421), and Positive epigenetic regulation of rRNA expression (R-HSA-5250913). Reactome also identified that, from this, 176 genes did not have any significant pathways or processes. A search was done to identify which genes were present in the top 25 pathways, and this resulted in 7 genes being identified, which were *EP300, MAPT, GRIA1, SNCA, GRIN2A*, and *PRKCA*.

Variant annotation:

A list of 1651 risk alleles associated with the final gene list containing 35 neurological traits was input into the Ensembl Biomart database. All the dataset counts are listed in Table 2. Doing this helped in the retrieval of information about the risk alleles, their associated gene variant, and details about how they contribute to their associated traits. Therefore, 29 attributes were selected in Biomart. The document produced from the results showed that there were 10,844 SNV variants associated in total with the risk allele data provided, with some risk alleles having multiple variant names, as shown in Figure 3. To collate information about the neurogenic effect of these SNV variants, we put the same list of rsIDs into the Variant Effect Predictor (VEP) tool of Biomart.

Table 2: Summary of dataset count.

Name	Count/Size
Initial dataset	1651 risk alleles
Final gene list	998 genes
Filtered model dataset (SNV)	11040 variants
Filtered model dataset (SV)	1173 variants

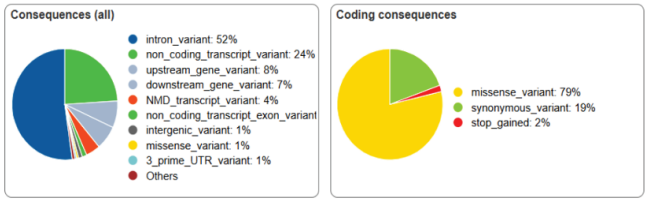

Figure 3: Visual summary of Variant Effect Predictor results. The figure explains that 1,746 SNV variants were processed from the input data set, with 1,381 overlapped genes and 7,805 overlapped transcripts between the risk allele IDs. (A) The first pie chart shows the percentage of the risk alleles associated with specific SNV variant types. (B) The second pie chart shows the percentage of the risk alleles that have an impact on the coding of proteins, categorized into 3 types: missense, synonymous, and stop-gained variant codon. From this, it is also understood that there were 1,737 existing SNV variants and around 9 novel variants identified, which can be used for further processing and analysis.

Following this, the information from Biomart and VEP was taken and collated into a single dataset. In this dataset all the unnecessary information was filtered out to a total of 19 parameters which included the rsIDs of the variants, variant location, allele, consequence, impact of variant, symbol, associated gene, feature type, bio type which specifies the variant function, existing variation, reference allele, uploaded allele, clinical significance, CADD PHRED score CADD RAW, REVEL score, clinical prediction, SIFT score prediction, Poly phenotype associated, label, and risk score. The target label here can be identified as the risk score. The final annotated variant sheet for SNV was then used in the classification model-building stage. Similarly, from the gnomAD database, the annotation of SVs was carried out.

SNV and SV Model evaluation:

The script was run for the SNV annotated dataset, where the mean CV, CV standard deviation, R^2, and MAE (Margin of Absolute Error) were calculated for 2 models - HuberRegressor and RidgeCV. Mean Absolute Error (MAE) is a measure of the average absolute error values of the predicted risk score and true risk score. These were then compared to find the best model, as seen in Table 2. The best model from this was RidgeCV with a CV R^2 value = 0.876; significantly higher than the HuberRegressor model. The algorithm also trained the 2 models to predict risk scores from the dataset, and this was compared against the true risk scores of the dataset. Figure 4 clearly shows how the RidgeCV performed during this training in the form of a visual scatter plot. The visualization of the results is important because it helps in understanding how the 2 models were trained, the accuracy of their predictions, and the comparison of their R^2 values.

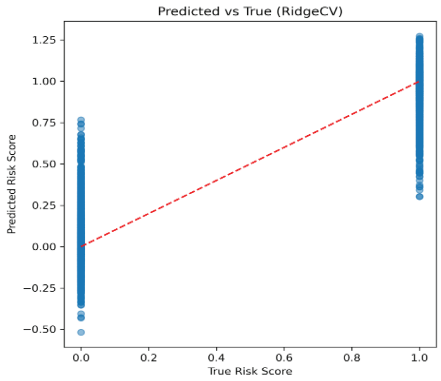

Figure 4: RidgeCV Predicted vs True Risk score scatter plot. The graph here shows how the model fared in prediction. In the graph, most of the points are concentrated at the extremes, which shows that the majority of the predicted scores were accurate to the true risk score of the dataset.

The algorithm also identified the most important features that affected the risk score prediction for the RidgeCV model. The top 5 features are Poly Phen prediction not available, Poly Phen prediction benign, CADD PHRED score, SIFT prediction tolerated, and SIFT prediction deleterious, low confidence. These predictors are relevant in predicting genetic evolutionary patterns by identifying the functional impact of mutations - changes in the base sequence of amino acids or DNA. They are essentially statistical tools that classify variants and help in understanding their genomic framework. On running the algorithm for the annotated variant data set for SV, similar to the SNV script, the mean CV, CV standard deviation, R^2, and MAE (Margin of Absolute Error) were calculated for 2 models, HuberRegressor and RidgeCV, as recorded in Table 3, and performed a quick comparison between the 2 models. From these results, it was inferred that RidgeCV was the best model, with a mean CV R^2 of 0.993. Both SV and SNV models were deployed into web applications using Streamlit. The web interface of both websites is shown in Figures 5 and 6 for SNV and SV, respectively.

Table 3: Summary of all models for SV.

Model	Mean CV	CV Std	R^2	MAE
HuberRegressor	0.89916	0.124901	0.930025	0.01547
RidgeCV	0.99317	0.012088	0.97954	0.002331

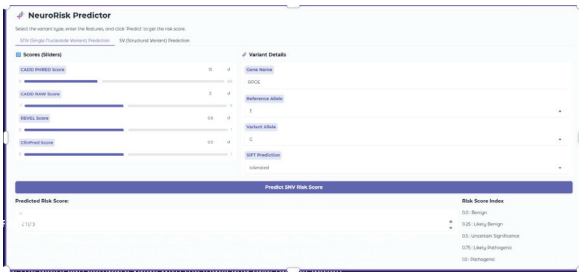

Figure 5: Single Nucleotide Variant (SNV) Risk Prediction Interface of the NeuroRisk Predictor. This interface presents the SNV (Single Nucleotide Variant) module of the NeuroRisk Predictor app. The left section allows users to input quantitative pathogenicity scores such as CADD PHRED, CADD RAW, REVEL, and ClinPred. The right section captures variant details, including Gene Name, Reference Allele, Variant Allele, and SIFT Prediction. The app primarily functions as a tool for research and support. It cannot work as a standalone clinical diagnosis tool, but in the future, this app can be developed further and can also be used as a monitoring system for users to keep track of their disease progression.

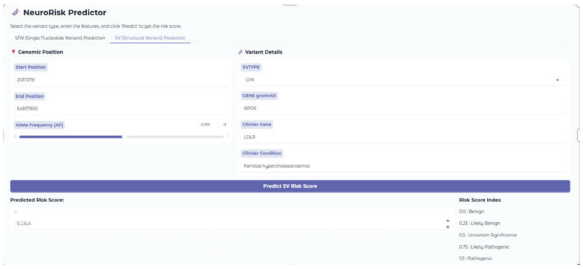

Figure 6: Structural Variant (SV) Risk Prediction Interface of the NeuroRisk Predictor. This interface displays the SV (Structural Variant) module of the NeuroRisk Predictor web application. Users input genomic coordinates (start and end positions), allele frequency (AF), and variant-specific annotations such as SVTYPE, GENE (gnomAD), ClinVar Gene, and ClinVar Condition.

Following this, the code provided a visualization of how the models were trained and the accuracy of the results in their prediction of genetic risk scores. Figure 7 shows the results for the RidgeCV model, where the predicted risk score of the model is plotted against the true risk score from the annotated variant data set.

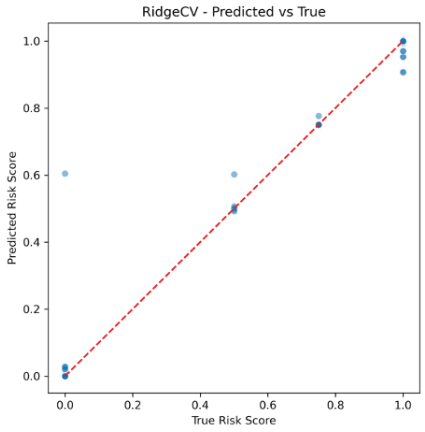

Figure 7: RidgeCV Predicted vs True machine learning scatter plot. The figure shows how the RidgeCV predicted risk score compared against the true risk score from the data set. As more dots are closer to the true risk score, the deviation from the actual values is also less. It can therefore be inferred that this model was more accurate in its prediction in comparison to the other model.

The algorithm then also identified which feature had the most impact on the risk score prediction for both models. It provided a visualization of the degree of influence of the features. The most important features for the RidgeCV model were neurological disorders, where the top 5 were Familial hypercholesterolemia, Retinitis pigmentosa, Familial aortopathy, Intellectual developmental disorder, and Nephronophthisis.

Discussion

The project began with a literature review of existing studies to gain an understanding of neurological disorders and how genome sequencing was done to identify target genes. Using the datasets from these studies, the genes acquired were analyzed in terms of their network with other genes and their neurogenetic functions. This helped identify the top contender genes responsible for these disorders. From this, the project tried to understand whether variants associated with these genes as SNVs and SVs were pathogenic and were the un-

DOI: 10.36838/IJHSR813.6

derlying genetic cause of these disorders. Following this, the project worked on annotating and understanding the variants associated with these genes. To build a machine learning model that could predict the pathogenicity and risk score of variants, the data collected was used to train 2 regression models - RidgeCV and HuberRegressor. These models were compared for accuracy, and the best one was found to be RidgeCV for both variant datasets. This model can aid in clinical and scientific analysis of associated variants of neurological disorders.

Pathogenicity prediction systems and classification systems of disease-causing variants are a crucial aspect for clinical interpretation and use. With the increasingly uncertain and rapidly changing nature of human genomic variants, having prediction systems that can be used to identify the risk score of a variant can also help in identifying patterns in the human genome. Studies have been developed on this idea for creating variant feature prediction systems for human genetic variation observed in large-scale populations.[14] This study specifically used variant calls from gnomAD and classified the variants by allele frequency for non-deleterious and deleterious rare variants. This data was annotated with the same feature used by CADD and was used as training data for their model. They trained logistic regression models, similar to this project, with the annotated variant datasets to predict the deleteriousness of the variant. They then used variants from the ClinVar database that were not used in the training to test the accuracy of their prediction system. Other studies have used the same idea but different machine learning models, such as neural networks, to predict the clinical impact of human variation.[15] This study has used deep neural networks to distinguish which mutations are pathogenic and which are benign by using common variants from non-human primate species that are identical in state to human variant sites. They used evaluation metrics such as accuracy and ROC curves on test sets and compared them against existing variant effect predictors. This study, in fact, proposed new candidate genes associated with intellectual disability based on their system and variant ranking.

This project has overlapping aspects with the other studies, but what stands out is the specification and focus on variants associated with neurological disorders. This project also explored in depth structural variants and single-nucleotide variants, rather than common variants, which avoid generalizability. The project has also found top genes responsible for neurological disorders after network analysis, and this could be useful for mapping genes and traits to the identified variants. The use of logistic regression models also provides transparent and interpretable coefficients, which, in comparison to other models like neural networks, are less complex. The models used in this project were also less prone to overfitting. They can be trained fast with a smaller dataset, therefore improving the speed and accuracy of the prediction system. However, the project could have implemented a larger training dataset for further development. The use of logistic regression, due to its more simplistic nature, may not capture non-linear interactions and model complex relationships, which are needed for effective understanding and accuracy of variant effect prediction systems. The project does not dive deeper into the phenotypic effects of these variants and could have been evaluated further with a training dataset from other databases for a thorough model evaluation. Future work on this project could include larger and more recent datasets. This would strengthen the ML model in terms of accuracy and validity. Additionally, the ML model would benefit from external validation with datasets from a different database than the one on which it was trained.

■ Conclusion

In summary, the findings from this project display how machine learning models like logistic regression can be highly effective in predicting pathogenicity of crucial variants in the form of a risk score. The exploration into and evaluation of structural variants and single-nucleotide variants demonstrate how deep an impact they have in causing neurological disorders. The increased efficiency and accuracy of prediction systems can, along with clinical interpretation, be beneficial in the classification and prioritization of disease-associated variants. While the focus on neurological disorders may limit the scope of the study, the ever-evolving nature of the genome can always be further analyzed. Future developments should integrate broader datasets to refine the scores and predictions further. These results, therefore, highlight how the enhanced nature of machine learning models and computational genomics can be used to develop fields of medicine and testing further, helping in the assistance of quick action to treat these disorders with the help of the identification of risk level and pathogenicity of the variants. This ML model and tool are not restricted to only neurological conditions. It can be trained and validated with datasets from other diseases to increase its scope and accessibility.

■ Acknowledgments

I would like to thank Ms. Aashna Saraf for all of her guidance and support throughout this entire process. I would also like to thank my parents and school, without whom none of this would have been possible.

■ References

1. Ehtezazi, T.; Sarker, S. D. Phytochemical Nanoparticles for the Treatment of Neurological Disorders. Phytochem. Anal. 2025, July. https://doi.org/10.1002/pca.70020.
2. New Global Action Plan on Epilepsy and Other Neurological Disorders Published. Accessed July 30, 2025.https://www.who.int/news/item/20-07-2023-new-global-action-plan-on-epilepsy-and-other-neurological-disorders-published?utm_source=chatgpt.com.
3. Website. https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(21)00164-9/fulltext (accessed 2025).
4. Chia, R.; Ray, A.; Shah, Z.; Ding, J.; Ruffo, P.; Fujita, M.; Menon, V.; *et al.* Genome Sequence Analyses Identify Novel Risk Loci for Multiple System Atrophy. Neuron 2024, 112 (13), 2142–2156.e5.
5. Akcesme, B.; Islam, N.; Lekic, D.; Cutuk, R.; Basovic, N. Analysis of Alzheimer's Disease Associated Deleterious Non-Synonymous Single Nucleotide Polymorphisms and Their Impacts on Protein Structure and Function by Performing In-Silico Methods. Neurogenetics 2024, 26 (1), 8.
6.. Al Khleifat, A.; Iacoangeli, A.; van Vugt, J. J. F. A.; Bowles, H.; Moisse, M.; Zwamborn, R. A. J.; van der Spek, R. A. A.; *et al.*

Structural Variation Analysis of 6,500 Whole Genome Sequences in Amyotrophic Lateral Sclerosis. NPJ Genom. Med. 2022, 7 (1), 8.
7. Erkal, B.; Akçeşme, B.; Çoban, A.; Vural Korkut, Ş. A Comprehensive In Silico Analysis of Multiple Sclerosis Related Non-Synonymous SNPs and Their Potential Effects on Protein Structure and Function. Mult. Scler. Relat. Disord. 2022, 68, 104253.
8. Ciulkinyte, A.; Mountford, H. S.; Fontanillas, P.; 23andMe Research Team; Bates, T. C.; Martin, N.
9. G.; Fisher, S. E.; Luciano, M. Genetic Neurodevelopmental Clustering and Dyslexia. Mol. Psychiatry 2025, 30 (1), 140–150.
10. Kaspi, A.; Hildebrand, M. S.; Jackson, V. E.; Braden, R.; van Reyk, O.; Howell, T.; Debono, S.; *et al.* Correction: Genetic Aetiologies for Childhood Speech Disorder: Novel Pathways Co-Expressed during Brain Development. Mol. Psychiatry 2023, 28 (4), 1664–1666.
11. Abe-Hatano, C.; Iida, A.; Kosugi, S.; Momozawa, Y.; Terao, C.; Ishikawa, K.; Okubo, M.; *et al.* Whole Genome Sequencing of 45 Japanese Patients with Intellectual Disability. Am. J. Med. Genet., Part A 2021, 185 (5), 1468–1480.
12. Dhaliwal, J.; Wagner, J. STR-Based Feature Extraction and Selection for Genetic Feature Discovery in Neurological Disease Genes. Sci. Rep. 2023, 13 (1), 2480.
13. Sinha, S.; Rabea, F.; Ramaswamy, S.; Chekroun, I.; El Naofal, M.; Jain, R.; Alfalasi, R.; *et al.* Long Read Sequencing Enhances Pathogenic and Novel Variation Discovery in Patients with Rare Diseases. Nat. Commun. 2025, 16 (1), 2500.
14. Nazaretyan, L.; Rentzsch, P.; Kircher, M. varCADD: Large Sets of Standing Genetic Variation Enable Genome-Wide Pathogenicity Prediction. Genome Med. 2025, 17, 84.
15. Sundaram, L.; Gao, H.; Padigepati, S. R.; McRae, J. F.; Li, Y.; Kosmicki, J. A.; Fritzilas, N.; *et al.* Predicting the Clinical Impact of Human Mutation with Deep Neural Networks. Nat. Genet. 2018, 50 (8), 1161.

Authors

Ishita Rao is an enthusiastic 11th-grade student from The International School, Bangalore, passionate about creativity and design. With a keen interest in problem-solving and innovation, she likes to explore hands-on learning through building, teamwork, and imaginative thinking.

Nirupma Singh is a Bioinformatics Scientist with a doctorate in Biotechnology and Bioinformatics from the University of Delhi, with six years of hands-on research and development experience. Her journey is marked by a robust foundation in Machine Learning and Python, with five years of expertise. Proficient in Linux and cloud servers like AWS. She excels in structural biology, systems biology, protein/gene network analysis, data mining, and computational genomics

DOI: 10.36838/IJHSR813.6