Predicting Parkinson's: Using Neural Networks to Evaluate the Genetic Risk Factors Behind the Disease

Andrew Yuan,[1] Isha Jagadish,[2] Trisha Gongalore[3] and Joseph Alzagatiti[#]

[1]Lynbrook High School, San Jose, CA, USA
[2]Saratoga High School, Saratoga, CA, USA
[3]Pomona College, Claremont, CA, USA
[#]Advisor

ABSTRACT

To date, researchers do not know the exact reasons for the loss of dopaminergic neurons in the substantia nigra pars compacta that leads to Parkinson's Disease (PD). Thus, it is extremely difficult to predict whether or not a patient is likely to develop the disease later on, as their risk increases with age. However, once patients present with the common symptoms indicative of the illness, a substantial amount of dopaminergic neurons are already lost. Seeing as there are no current avenues of replacing those neurons, predictive diagnosis and preventive measures could be of extraordinary help in devising treatments. Our aim was to use the research into possible high-risk genetic factors from genome-wide association studies (GWAS) to formulate a predictive neural network model for Parkinson's. We analyzed patient genomes for mutations in the top 20 genes associated with PD, as well as 21 genes implicated in axon guidance pathways, to determine whether the patients were at high or low risk for Parkinson's. Our model produced an accuracy and AUROC of 94%. We found this significant because it showed a strong correlation between the single nucleotide polymorphisms (SNPs) we analyzed and PD. We believe our model can be further improved upon by adding considerations for other investigated risk factors, such as patient age, familial history of disease, or gut microbiota inconsistencies among others.

Introduction

Parkinson's Disease stands as the second most common neurodegenerative disease with an average onset of 70 years old[1]. The disease is characterized by a loss of neurons that produce dopamine in the substantia nigra pars compacta, a basal ganglia structure in the midbrain. Research has shown that this neuronal cell loss is in large part due to the accumulation of misfolded, phosphorylated α-synuclein proteins, forming aggregates called Lewy Bodies[2]. PD patients display significant motor deficiency symptoms, including difficulty maintaining balance, freezing of gait, and difficulty initiating movements like walking. Tremors, which worsen with the progression of the disease, bradykinesia, and muscle rigidity are other symptoms. Non-motor symptoms consist of dementia, sleep disorders, and depression[3]. Current treatments include administering drugs like Levodopa (L-Dopa), a molecule similar in structure to dopamine, in the striatum to facilitate increased dopamine uptake, as well as deep brain stimulators that are surgically implanted and stimulate the motor cortex to rescue loss of motor function[4].

However, by the time patients present with symptoms, they have already been found to exhibit Lewy Body pathology and neuronal cell loss[5]. Consequently, scientists have been conducting research into the causal factors of the disease in an effort to prevent neurodegeneration in the first place. One avenue of study has been into the genetic factors because of the way certain genes have recently been found to be correlated to certain diseases (for example, the BRCA1 gene and breast cancer)[6]. Parkinson's occurs in two forms: the familial form, which accounts for around 10% of cases, and the sporadic form, which accounts for around 90% of cases[7]. Several GWAS have been done to take a deeper look into the genetic risk factors associated with both familial and sporadic PD, identifying mutations in the SNCA, LRRK2, PINK1, and PARK genes among others that have been shown to increase the likelihood of contracting the disease[8]. The study conducted by Maraganore et al. 2005 identified many single nucleotide polymorphisms (SNPs), the substitutions of a single base in a DNA sequence, present in each patient's genome and analyzed them to determine which ones were most prevalent in individuals with PD and which ones might constitute a higher risk factor[9]. In a subsequent study, several of these SNPs were found to be involved in axon guidance pathways, or the processes by which axons make connections with other neurons. The failure of these pathways has been implicated in increasing PD risk in patients[10].

Our research strives to utilize deep learning neural networks, a subset of machine learning, to explore the influence of multiple genetic risk factors on the likelihood of a patient contracting Parkinson's Disease. Creating deep learning models occurs in several stages: the programming stage, training stage, and the ready-for-use stage. In the programming stage, engineers encode specific algorithms into the deep neural network that are relevant to the information that will be processed. They then train the neural network on large amounts of data, in which the program will be able to identify patterns to predict the target outcome in future cases. This is the stage where neural networks differ from machine learning, in that there is less preprocessing of data required before beginning training because the model does not need help identifying the importance of different patterns it finds: it will be able to determine that on its own. Finally, engineers can input data, and based on the patterns identified in the training stage, the program will generate outputs. In the case of this model, the completed neural network was able to take in a patient's genetic data as the input to the model and output a prediction for whether the patient had PD (either high risk or low risk).

Review of Literature

In the past, machine learning has been used to determine the severity of the phenotype displayed by the diseased individual. For example, one algorithm utilized videos of gait, drawing patterns based on the way patients walked and focusing especially on the freezing of gait that is particularly significant in Parkinson's patients[11]. Another model was trained to analyze differences in the way Parkinson's patients write or draw compared to healthy controls due to the fact that Parkinson's individuals have resting tremors that disrupt their ability to write or draw with a steady hand[12]. There are even algorithms that analyze voice recordings to look for patterns in speech that vary from Parkinson's individuals to healthy individuals[13]. However, there hasn't been as much research into how the genetic components of Parkinson's can be used to develop a predictive model; hence, that was the direction our research took.

Methods

Data Sources

We retrieved patient genome data from Phase 3 of the International HapMap Project[13], which included patients from 11 different populations (ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, CHD: Chinese in Metropolitan Denver, Colorado, GIH: Gujarati Indians in Houston, Texas, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MXL: Mexican ancestry in Los Angeles, California, MKK: Maasai in Kinyawa, Kenya, TSI: Toscani in Italia, and YRI: Yoruba in Ibadan, Nigeria). As for gene data, we looked at Text File 1[14] located in the Supplementary Materials section from the GWAS conducted by Maraganore et al. 2005[8]. Furthermore, we utilized the results of the paper by Lesnick et al. 2007[14], which sought to investigate the correlation between genes related to axon-guidance pathways and their effect on PD. Specifically, we used the results from Table 1 in their paper (Supplementary Materials #4), which listed all the SNPs in genes expressed in axon-guidance pathways in the brain that were proven to have a higher correlation to PD, to help us construct our neural network.

Data Preparation

Before constructing the neural network model, we had to prepare the data. All of our work was implemented in Python and was conducted in a Google Colab Notebook, which is essentially an environment that supports libraries such as Pandas, Keras, and TensorFlow, and allows one to build and train a machine learning model.

Processing the Gene Data

We loaded the SNPs related to axon-guidance pathways and then cropped this dataframe to only include the rsID's (specific way to characterize each gene mutation), as all other columns were irrelevant for this data processing stage. We also loaded the GWAS SNPs from the aforementioned Text File 1. However, the amount of SNPs in the GWAS data was approximately 198,000, so we decided to only include the 20 SNPs with the lowest P-values (indicative of the highest correlation to PD). Next, we cropped this dataframe to only include the rsID's. Finally, we combined both the 20 GWAS rsID's and the 21 axon-guidance pathway rsID's to obtain an array of the rsID's of all 41 SNPs that were proven to have the highest correlation to PD (definitive PD SNPs) (Table 1).

Table 1: Definitive PD SNP and Gene Function Table. This table includes the 20 SNPs found to be most correlated with Parkinson's Disease based on the patient data from the Maraganore et al. GWAS study, as well as the 21 SNPs found to be associated with axon guidance pathways as an extension of the same study. The table further expands on the gene in which the SNP is located as well as the general function of the gene. See Supplementary Materials for full table.

Definitive PD SNP and Gene Function Table		
rsID	Gene Name	Function
rs10815285	ERMP1	activates endoplasmic reticulum metallopeptidase activity
rs10917325	EPHB2	encodes receptor tyrosine kinase transmembrane glycoproteins, responsible for division and motility
		● ● ●
rs9688032	SLIT3	involved in effecting cell migration
rs9789345	SLC8A1	encodes solute carrier family 8 member A1

Processing the Patient Data

Next, we had to load each of the 11 patient data files for each population. We kept only the SNPs for each patient which were present in the aforementioned definitive PD SNPs. We then computed the percentage of definitive PD SNPs that were present in each patient by dividing the count of definitive PD SNPs present by the total number of definitive PD SNPs (41). Finally, we had to make an educated guess as to whether each patient was at high-risk or low-risk for contracting PD. If the calculated mutation percentage was greater than the threshold value of 30%, we annotated the patient with a "1" (high-risk), and if the percentage was less than or equal to 30%, the patient was marked with a "0" (low-risk).

Constructing the Neural Network

Our goal was to construct a supervised machine learning model where the input was the patient SNP data and the output was the model's prediction of the patient's risk level (high or low).

Preparing Training/Test Datasets

In order to construct well distributed training and test subsets from the curated patient data, we used the following approach:

- First, we evaluated each population in the patient data set by further grouping them into high-risk and low-risk groups based on the risk value (0 or 1).
- Then for each risk group, we further divided the high-risk group into a 90/10 train-test split and the low-risk group into a 40/60 train-test split.
- Finally, we aggregated all of the training and test subgroups across all 11 populations into combined training and test datasets.
- We also prepared a second group of 11 test datasets corresponding to each of the 11 populations, so that population-wise testing could be done (Fig. 4).
- We then compared the accuracy produced by using the test subgroup across all 11 populations and the accuracy produced by testing individual populations as a whole.

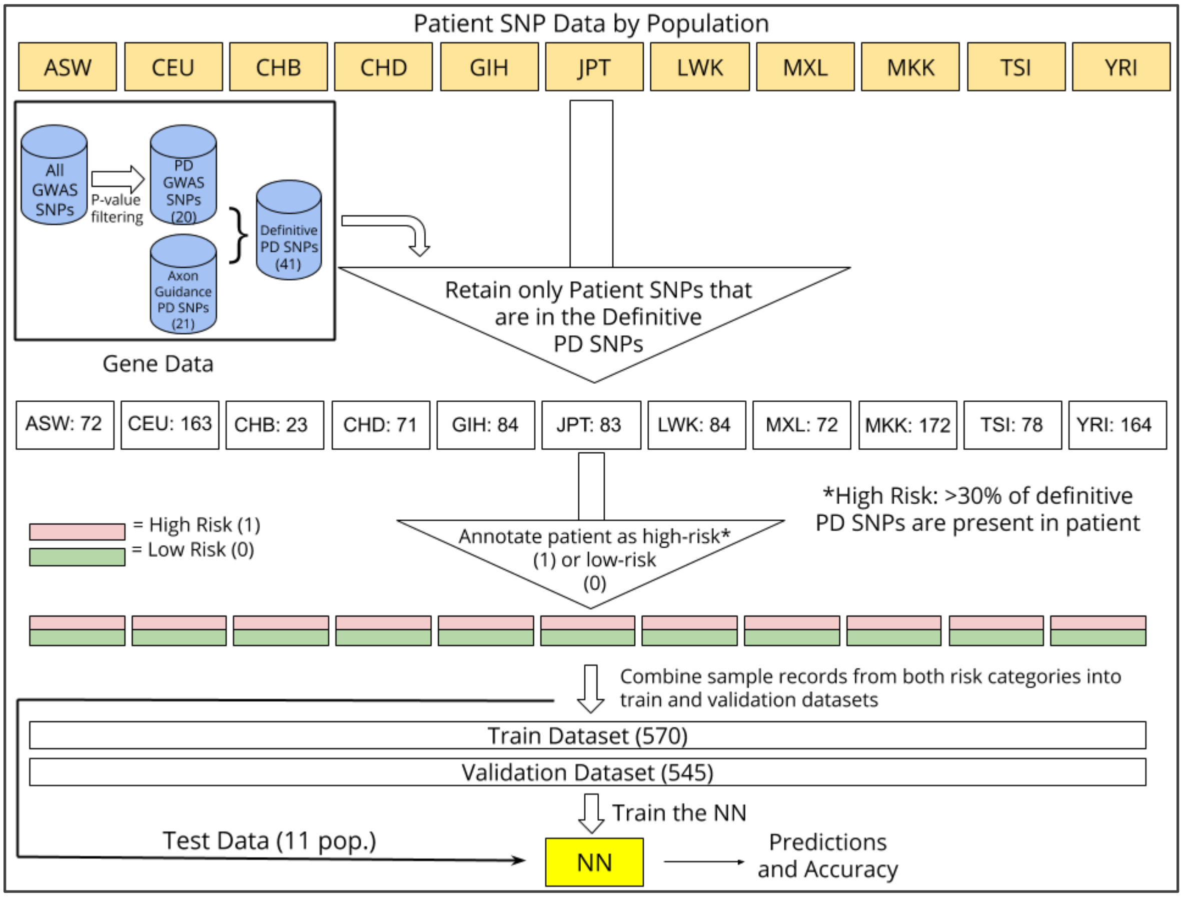

Figure 1. This flowchart illustrates the steps we took in order to conduct our research. The first step was the data preprocessing, which is highlighted in blue and yellow. Finally, the model outputs its prediction of a patient's risk value, as well as the scores of various metrics that were used to evaluate the model's performance.

Neural Network Design/Execution

We constructed a multilayer perceptron (Deep Neural Network) architecture to compute whether a patient was at high-risk or low-risk for PD. Our model was Sequential and consisted of one input Dense layer, multiple hidden Dense layers, and one output Dense layer. The input and hidden layers had multiple nodes (neurons) while the output layer had only two nodes, corresponding to the two categorical outputs (0 or 1). We used ReLU as the activation function for the input and hidden layers and Softmax as the activation function for the output layer (Fig. 2). Additionally, we utilized a Dropout layer in between the Dense layers with a Dropout rate of 0.25 to prevent overfitting. Finally, when compiling our model we used the 'Adam' optimizer and a 'Binary Cross-Entropy' loss function.

We experimented with multiple architectures by modifying the number of layers and nodes per layer to achieve optimal accuracy. For each architecture, we trained our model on the training dataset. We set the number of Epochs to 50 during each run. We then tested that model with each of the individual population datasets and compared performances using various metrics (see Results for analysis of different metrics).

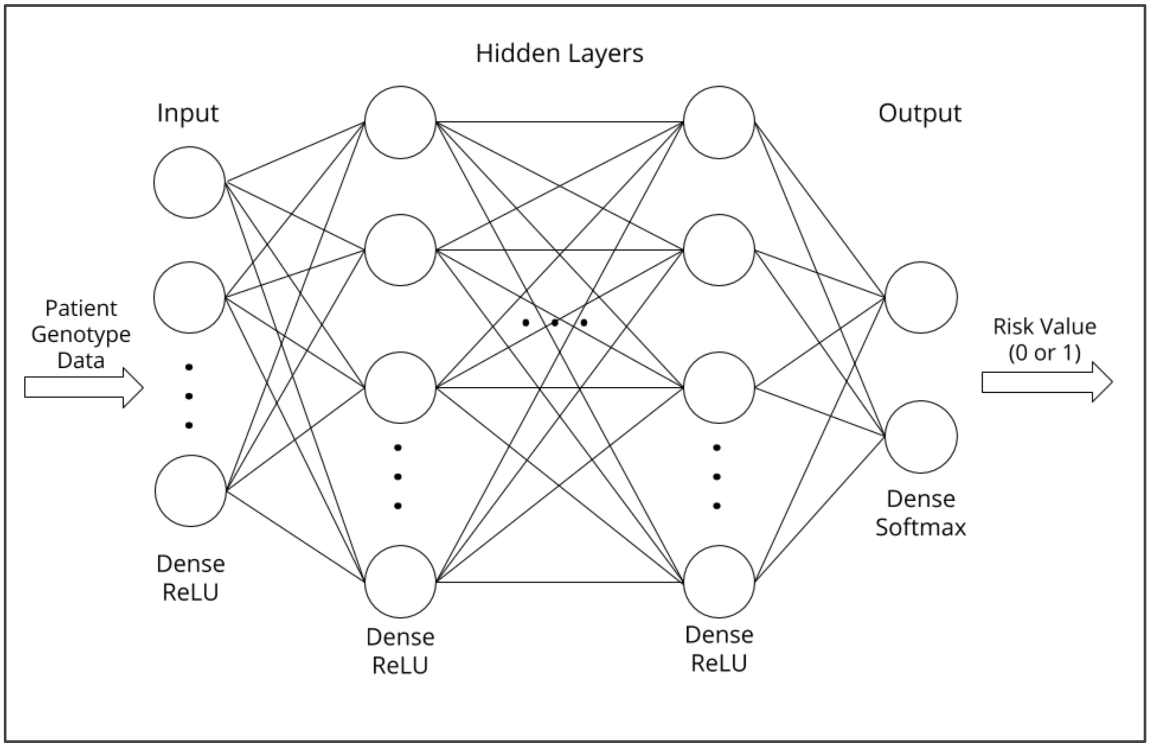

Figure 2. This figure details the basic outline of the deep neural network we constructed. The input is the patient genotype data while the output is the patient's risk of contracting PD, classified as a 0 or 1.

Results

Various Methods to Evaluate Neural Network Performance

When evaluating the performance of neural networks, there are many tests one may use. The simplest and most popular test is the accuracy test.

Another test we carried out was the F1 score, which is the harmonic mean of precision ((TP)/(TP+FP)) and recall ((TP)/(TP+FN)), where TP, FP, and FN is the number of true positives, false positives, and false negatives respectively. An increase in precision often leads to a sacrifice in recall, and vice versa, so the precision-recall tradeoff is averaged into a reliable F1 score.

The Area Under the Receiver Operating Characteristic Curve (AUROC) is the final score we computed. The Receiver Operating Characteristic plots the true-positive rate against the false-positive rate, and is also not affected much by imbalances in the data.

Comparison of Different Neural Network Architectures

One of the areas in which neural networks differ the most is architecture. After testing many different architectures with different numbers of layers as well as different numbers of nodes in each layer, we found that our highest-performing neural network, 41-20-2, achieved an accuracy of 94%, an F1 score of 70%, a Cohen's Kappa of 68%, and an AUROC of 94% (Fig. 3). Because accuracy neglects the imbalances in the data, we used F1 score, Cohen's Kappa score, and AUROC as more reliable ways for comparing the different architectures.

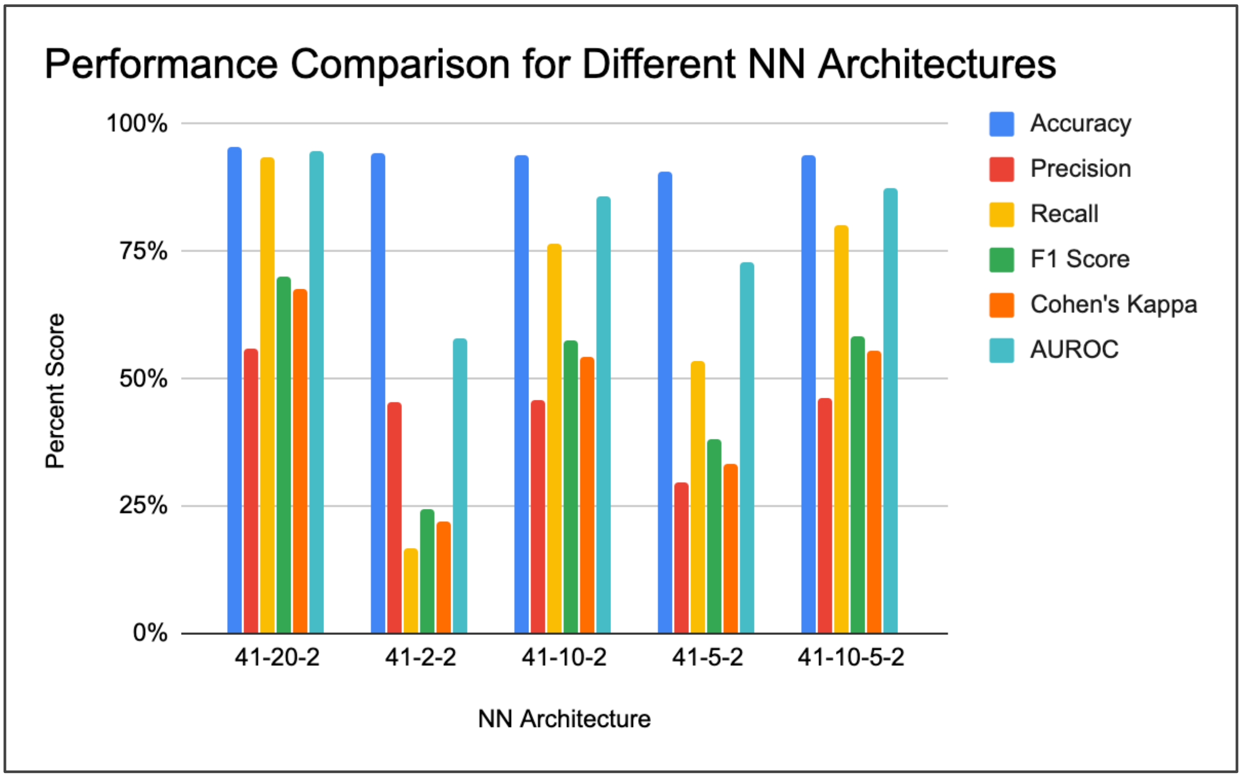

Figure 3. This graph compares the scores of various metrics that were used to evaluate our model's performance when testing each population as a whole.

Accuracy for Data Sets of Varied Composition

The population our best model (41-20-2) performed the best on was TSI (Toscani in Italia) with an accuracy of 96%. The population our model performed the worst on was GIH, (Gujarati Indians in Houston, Texas) with an accuracy of 81%. In testing each individual population, we achieved an average accuracy of 88%, compared to our overall testing accuracy of 95% (Fig. 4).

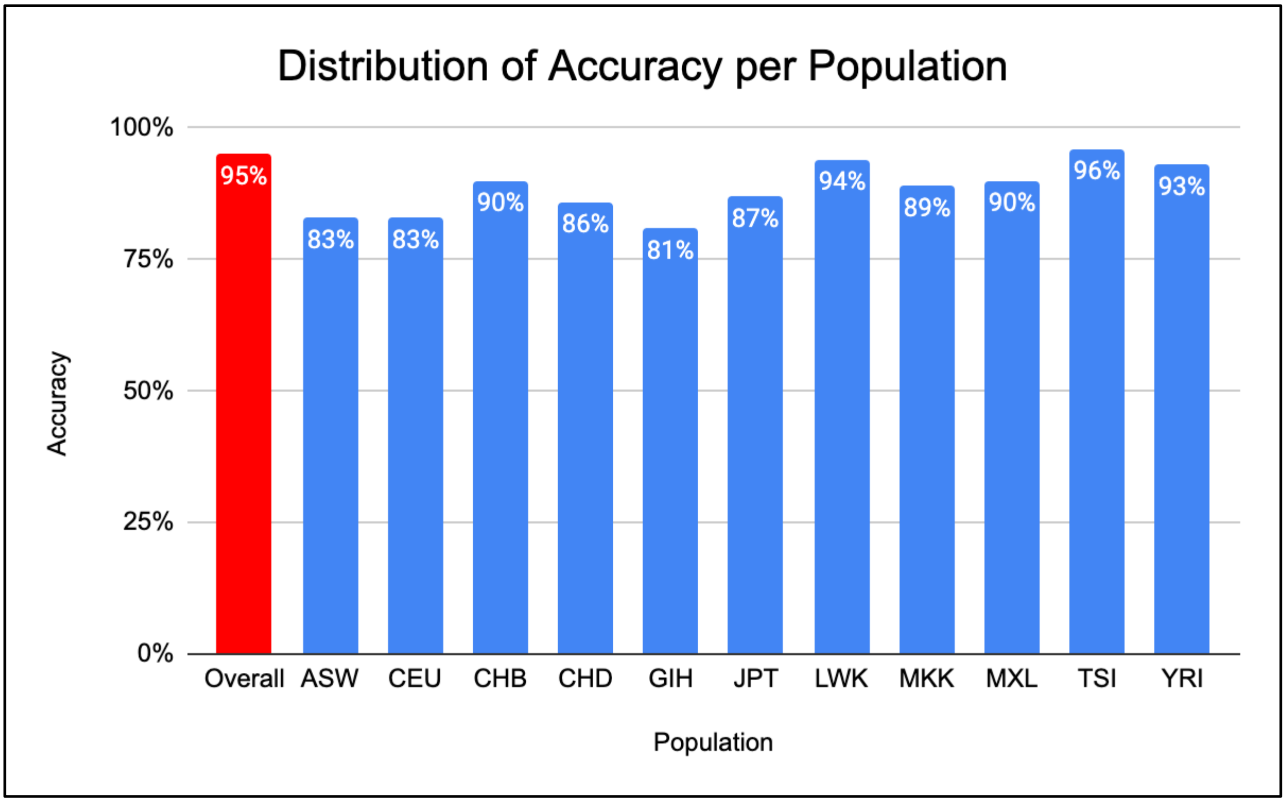

Figure 4. The graph above shows the distribution of accuracy scores across the 11 different population groups (shown in blue), as well as the accuracy of our model when testing against all 11 populations overall (shown in red).

Discussion

A drawback of the accuracy test is that it does not account for the imbalances in the training and testing data, which spawn from an uneven distribution of the high-risk patients and low-risk patients in the data. One test we used that resolves this problem is the Cohen's Kappa coefficient, which is a measure of the agreement between the model's predictions and actual outputs compared to pure chance. However, the Cohen's Kappa, F1 score, and AUROC, unlike accuracy, are still very reliable when the data is greatly skewed. When discussing Cohen's Kappa scores, a Kappa coefficient above 0.60 is generally accepted as a "good" result[15], meaning our model has significant predicting power.

Conclusion and Implications

Ultimately, we sought to create a neural network model that learned to correlate 41 different SNPs to whether a patient was high-risk or low-risk to PD. Currently, there is no neural network model that exists with the ability to determine a patient's risk to PD with high accuracy and convenience, making our findings extremely relevant. Our model can be used in the real world to predict a patient's risk to PD, simply given their genotype sample.

Future Research

We seek to expand this model to be able to perform a higher complexity, non-binary task, such as multi-class classification into different risk groups or outputting percentage risks for PD while maintaining high accuracy. Furthermore, more genes or other risk factors, like patient age and prodromal illnesses, may be utilized in future models. Instead of a supervised neural network, an unsupervised machine learning algorithm could be used to group patients into risk groups. In the future, our model can be significantly improved if it can be trained on a combination of patient genotypic and phenotypic data, such as additional SNPs, voice patterns, drawing samples, and gait analysis, resulting in a more accurate PD predictor.

Limitations

We were able to achieve our high model accuracies largely due to the simple task complexity, as our model only needed to compute a binary classification. One challenge we encountered was that in the beginning, our model predicted almost all low-risk, and was extremely hesitant to predict high-risk. We hypothesized that this was due to the fact that our training data was highly imbalanced, having significantly more healthy patients than PD patients. Because of this, our neural network algorithm quickly learned that only predicting low-risk would easily achieve high accuracy. To fix this, we adjusted the test-train split from 80/20 to 90/10 for the high-risks and from 80/20 to 40/60 for the low-risks. After this rectification, our model predicted fewer false negatives, improving our overall accuracy.

Acknowledgments

We would like to graciously thank Dr. Lina Kim and all of the Science Research Academy staff at UCSB for making our experience a very pleasant one. Most importantly, we would like to thank our mentor Joseph Alzagatiti and our teaching assistant Ian Taylor for helping us write our research paper and providing us with support throughout our journey at SRA.

References

1. NINDS. (2020) Parkinson's Disease: Hope Through Research. *NIH Publication* 20,139. https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Hope-Through-Research/Parkinsons-Disease-Hope-Through-Research#whatis
2. Shahmoradian, S.H., Lewis, A.J., Genoud, C. *et al.* (2019) Lewy pathology in Parkinson's disease consists of crowded organelles and lipid membranes. *Nat Neurosci* 22, 1099–1109. https://doi.org/10.1038/s41593-019-0423-2
3. Chaudhuri, R., Healy, D., Schapira, A. (2006) Non-motor symptoms of Parkinson's disease: diagnosis and management. *The Lancet Neurology* 5 (3), 235-245. https://www.sciencedirect.com/science/article/abs/pii/S1474442206703738
4. Rascol, O., Payoux, P., Ory, F., Ferreira, J., Brefel-Courbon, C., Montastruc, J. (2003) Limitations of current Parkinson's disease therapy. *Annals of Neurology* 53 (S3). https://doi.org/10.1002/ana.10513
5. Postuma, R., Berg, D. (2016). Advances in markers of prodromal Parkinson disease. *Nat Rev Neurol* 12, 622–634. https://doi.org/10.1038/nrneurol.2016.152
6. Hall, J., Lee, M., Newman, B., Morrow, J., Anderson, L., Huey, B., King, M. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 250 (4988): 1684–1689. doi:10.1126/science.2270482
7. Thomas, B., Beal, M. (2007) Parkinson's disease. *Human Molecular Genetics* 16 (R2), 183–194. https://doi.org/10.1093/hmg/ddm159

8. Maraganore, D., Andrade M., Lesnick, T., Frazer, K., Cox, D., Ballinger, D. et al. (2005) High-Resolution Whole-Genome Association Study of Parkinson Disease. *American Journal of Human Genetics* 77, 685-693. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1271381/#ui-ncbiinpagenav-heading-3
9. Livesey FJ, Hunt SP (1997) Netrin and netrin receptor expression in the embryonic mammalian nervous system suggests roles in retinal, striatal, nigral, and cerebellar development. Mol Cell Neurosci 8, 417–429. https://doi.org/10.1006/mcne.1997.0598
10. Berus, L., Klancnik, S., Brezocnik, M., & Ficko, M. (2018). Classifying Parkinson's Disease Based on Acoustic Measures Using Artificial Neural Networks. *Sensors (Basel, Switzerland)*, *19*(1), 16. https://doi.org/10.3390/s19010016
11. Gil-Martín, M., Montero, J., San-Segundo, R. (2019) Parkinson's Disease Detection from Drawing Movements Using Convolutional Neural Networks. *MDPI electronics* 8(8), 907. https://doi.org/10.3390/electronics8080907
12. Maachi, I., Bilodeau, G., Bouachir, W., (2019) Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait, *Cornell arXiv.* arXiv:1910.11509
13. Team, W. (2002). *HapMap 3*. Wellcome Sanger Institute. ftp.ncbi.nlm.nih.gov/hapmap/phase_3/hapmap3_reformatted/
14. Lesnick, T., Papapetropoulos, S., Mash, D., French-Mullen, J., Shehadeh, L., de Andrade, M., et al. (2007) A Genomic Pathway Approach to a Complex Disease: Axon Guidance and Parkinson Disease. *PLoS Genet* 3(6), 98. https://doi.org/10.1371/journal.pgen.0030098
15. Landis, R., Koch, G. (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159-174. https://pubmed.ncbi.nlm.nih.gov/843571/

Supplementary Materials

1. Table 1: Definitive PD SNP and Gene Function Table: https://drive.google.com/file/d/1Kw2aTMgCEOgC1BF4ndjyCVNaLV19lv74/view?usp=sharing
2. Patient Data (HapMap Project Phase 3): ftp.ncbi.nlm.nih.gov/hapmap/phase_3/hapmap3_reformatted/
3. GWAS Gene Data:
    a. Text File 1: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1271381/bin/AJHGv77p685tableS2new.txt
    b. Text File 2: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1271381/bin/AJHGv77p685tableS3new.txt
4. SNPs in Axon-Guidance Pathway Genes: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1904362/table/pgen-0030098-t001/