HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Development and Testing of Machine Learning Applications into Genetic-Based Disease Detection

Aditya Mittal[1] and Kuan-Chen Wu[#]

Affiliation
[#]Advisor

ABSTRACT

This study uses publicly available gene-expression from peripheral blood mononuclear cells fed into a logistically trained machine learn model to accurately predict the probability of early onset of Multiple Sclerosis by identifying biomarkers in genetic expression and establishes logistic regression as a viable methodology for genetic analysis to predict disease. Current detection methodology of neurological diseases such as MRI scans of existing lesions are impractical solutions when it comes to alleviating most of a patient's symptoms, as they rely on the disease to have already developed to detect it. Machine learning is a rapidly emerging tool that has much potential in not only disease detection, but early onset diagnosis as well. This study utilized the NEO Gene Expression Omnibus data repository to selectively identify key PBMC gene expression datasets to feed into a logistically trained model. Data filtration by Log-Fold Change analysis and p-Value importance allowed for data simplification to reduce model dimensionality, improve model accuracy, and even identify important gene markers in Multiple Sclerosis. Nearly 33,000 genes were eliminated through extensive data filtration, and 15 genes were marked as statistically significant in the development of Multiple Sclerosis. Model accuracy produced was nearly 100%, though lack of representative data highlights the need for further testing. The methodology in this experiment from the data accumulation to the actual construction and testing of the model itself serves as strong representation of the value artificial intelligence can have in the field of genomic analysis in disease detection.

Introduction

Affecting more than 2.8 million people in the world today, MS is the most common immune related disease affecting the central nervous system, which comprises of both the brain and spinal cord [1]. MS is a demyelinating disorder which causes the insulating covers of the nerve cells both in the brain and spinal cord to be eroded away, causing a variety of physical, mental, and even psychiatric problems. As seen in Figure 1, MS presents a disparity between healthy and diseased sections of the brain, as demyelination strips affected areas of almost all its mass, rupturing the vital passage of communication through which the brain sends signals to the rest of the body.

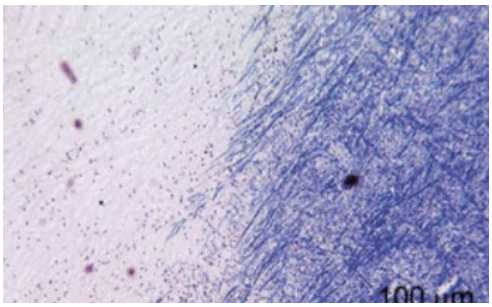

Figure 1. Distinct separation between demyelinated neurons (light blue) and healthy neurons (dark blue) [2]

The popular method of MS diagnosis today is by analyzing Magnetic Resonance Imaging (MRI) data to detect the number and volume of existing lesions caused by the MS in the central nervous system [3]. However, diagnosis via MRI remains difficult and error prone. Recent advances in ML techniques applied to MRI analysis has shown promising results, but identification of phase of the disease and disease prognosis remains poor [4]. Most importantly, since this method depends on lesions already caused by MS, detection is usually in later phases of the disease, when it is less treatable.

Therefore, genetic detection of this disease can have significant impact on early detection of MS and subsequent avoidance of its severe impact. While MS is not a hereditary disease, genetic variations have been shown to be associated with the disease [5]. Studies have shown that peripheral blood mononuclear cells (PBMCs) bear specific dysregulations in genes and pathways at distinct stages of MS that may help with classifying MS and non-MS subjects, identifying the early stage of disease [6]. More specifically, the differential expression of small microRNA (or miRNA) in PBMC plays a key role in MS because miRNA profiles are altered within CNS lesions and in the immune system and affect gene expression in many cell types involved in the disease [7]. To further: "A global consideration of miRNA dysregulation and the resulting alterations in gene expression may provide valuable insights into the pathophysiology of MS and reveal new alternatives for early diagnosis and treatment [8]." The combination of miRNAs' importance in gene expression in MS and its stability for scientific research makes it an asset when studying to identify biomarkers in MS. Available repositories of miRNA gene expression in peripheral blood mononuclear cells serve as the main source of the data, and cross-referencing expression values led to an accurate and precise dataset.

While a large amount of gene expression data, containing various biological implications, exists, the challenge is to detect a panel of discriminative genes associated with MS. This is where machine learning can play a huge role in mining large quantities of data to build a predictive model and also assign importance to specific features of the data set which ultimately help to isolate key genes that play an important role in the formation of MS. These advantages, among others, have made neural networks increasingly important in the field of genomics, and a goal of this study is to further the usage and highlight the importance of machine learning in science.

In this study, an unbiased machine learning workflow to identify MS classifiers using gene expression data found via PBMC cells is described. In addition, an analysis to identify disease-related genes using multiple sclerosis as an example is performed. The outcome of this project is to build a neural network that can output the probability of developing early onset MS through miRNA data from PBMC Microarrays to find significance of gene expression in MS. Through the utilization of online repositories of gene expression data and through the construction of an Artificial Neural Network, it is reported that machine learning is an extremely viable and accurate method to detect disease and has the potential to be implemented in real world applications. This research to identify biomarkers yielded specific importance around the MALAT1 and SAMSN1 genes, which can be further studied to reveal crucial information regarding the pathology of MS.

While this specific project focused on the neurodegenerative disease Multiple Sclerosis, a fundamental goal of this research is to develop a generic methodology that can comb through vast genetic data in an efficient and accurate way to diagnose or even predict a specific disease ahead of time.

Methods

Data Collection – Microarray GEO Database Selection

A fundamental principle of this study is to utilize machine learning methods to identify MS in a patient long before physical indicators have formed, as at that point the disease is well beyond curable. As shown in Figure 2, Machine Learning in MS can be used to process raw images such as MRIs or analyze microarray data. The problem with Image processing is that the disease is then treated as damage control – no real solutions can be implemented, and doctors and victims alike are on the back foot, simply responding to sudden spikes in the disease itself.

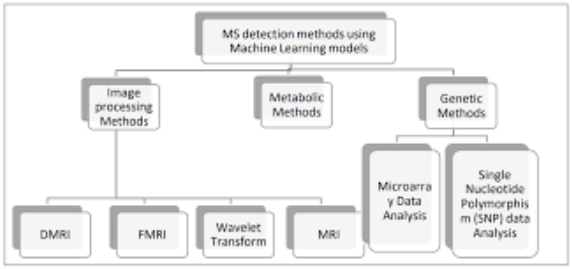

Figure 2. Differing ways machine learning can be used in MS detection models – Image processing is often seen as a poor choice as it can only be effective once the disease has already manifested itself [10].

To create something that is truly beneficial, a model that can preemptively determine MS before the disease has even formed itself is essential. Gene microarray analysis is the strongest method in order to achieve this goal. Microarrays used for gene expression is a rapidly advancing technology, allowing researchers to identify genes that are expressed in the sample with utmost accuracy [10].

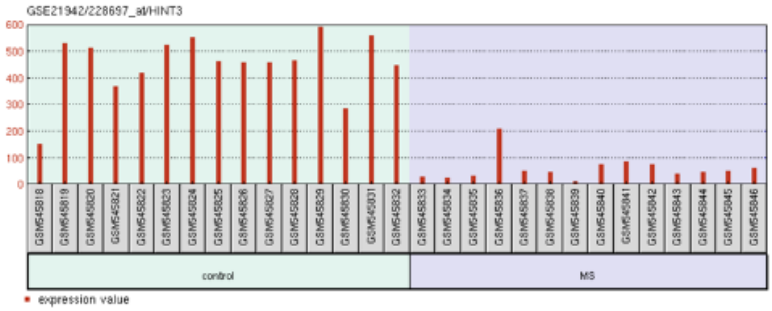

Figure 3. The HINT3 gene expression values were particularly different between healthy and diseased samples – a strong indicator the HINT3 expression plays some role in MS.

The gene data used for this study were accessed from the Gene Expression Omnibus Repository (GEO). The two datasets, accessed via the GEO repository, used in this study both utilized microarrays to analyze mRNA samples to record levels of expression as they reviewed the levels of miRNA expression in peripheral mononuclear blood cells. Peripheral mononuclear blood cells play an important role in MS as established, so the studies used provide the perfect data to analyze [2]. Specifically, both studies used the CHOP

Human Gene 1.0 ST array to compare specific expression profiles of genes found in healthy controls and multiple sclerosis patients.

The first database (GEO – GSE23832) containing 4 healthy controls and 8 diseased samples were combined with the second dataset (GEO – GSE21942), which contained 15 control samples and 12 diseased samples to create a much larger and more usable dataset.

In total, a matrix of 31 features seen over 27 samples is isolated from the original data.

To perform an initial analysis of the data, the GEO2R Analysis tool provided by the GEO was used to identify a multitude of important statistics including Log Fold Change in expression, along with p-values. For each gene, a graph displaying the differing levels of expression in the control and diseased patient was produced. For example, Figure 3 demonstrates a sharp disparity between the levels of expression for the HINT3 gene in a control patient compared to a patient suffering from MS. Viewing these differences numerically serves as a key data point that can be used in a neural network.

Data Filtration – LogFC Screening

A major problem with the data once retrieved from the GEO is its large size. Over 54,000 genes are analyzed per study, with over 85% of them being interevent to the project as a whole. This excess data serves a major problem for machine learning and logistic regression functions. The large number of features to analyze combined with the small amount of training and testing data available makes data filtration mandatory to develop a successful network.

The primary tool used for data filtration in this study was LogFC value comparison. LogFC values present the average fold change, or difference, in gene expression of a single gene between all the healthy and diseased datasets. A high positive LogFC indicates a large upregulation of gene expression in Multiple Sclerosis patients, while a large negative LogFC indicates a large downregulation of gene expression in Multiple Sclerosis patients. In order to filter down the data into manageable sizes for the neural network, genes were first ranked in order of LogFC gradience. Next, all genes with LogFC values greater than 2 and less than -3 were identified as statistically significant. This measurement was key in filtering a large number of genes in the datasets. While the GEO datasets originally contained more than 55,000 genes, after the LogFC screening, only 30 genes were identified as both statistically and biologically important. This not only serves as the first step towards a machine learning algorithm, but this methodology in itself also already identifies key genes in patients as statistically significant to MS.

Logistic Regression Implementation

Now that the data has been created and isolated, a logistic regression model can be created. Logistic regression serves as the primary model used to classify between multiple sclerosis and healthy. Logistic regression is an important tool in artificial classification, and it a simple neural network that can be used to differentiate between spam and regular emails, or in this case, healthy versus potentially diseased gene datasets.

As seen in Figure 4, the various libraries imported make up the logistic function, and allow our data to be read. The Pandas library is used to extract the data from the raw csv file placed in the computer's directory. The matplotlib and seaborn functions can be used to plot specific gene expressions and can be used to visually represent feature importance.

```
In [67]: #import basic libraries
         import pandas as pd
         import numpy as np

In [68]: #import plotting libraries
         import matplotlib.pyplot as plt
         import seaborn as sns
         %matplotlib inline

In [69]: #Data Gathered from GEO
         data = pd.read_csv("msData.csv")
```

Figure 4. Basic libraries imported for various purposes. Data is extracted via csv file.

Once all the required libraries are imported along with the data, it is crucial to split the data from the labels. In Machine Learning, the difference between supervised and unsupervised learning is the addition of labels along with data. In this study, the data is coupled with a label which indicates whether the gene expression values lead to a MS patient or a healthy one. In order to feed just the data and not the label associated with it, a drop is needed as shown in Figure 5.

```
In [70]: #Isolating data - Seperating Label
         X = data.drop("Outcome", axis = 1)
         y = data["Outcome"]
```

Figure 5. Data is separated from its label of diseased (1) or healthy (0).

Once the data has been entered, the program utilizes the Scikit-Learn library as the repository for its machine learning model. As seen in Figure 6, both splitting data and classification of the data can be used via the Scikit-Learn library. Once the data has been split into training and testing it can be fit into the actual model imported from the Scikit-Learn library.

```
In [71]: #import random split
         from sklearn.model_selection import train_test_split

In [72]: #split data
         X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.3, random_state=101

In [73]: #import model
         from sklearn.linear_model import LogisticRegression
         logmodel = LogisticRegression()

In [74]: #fit data into model
         logmodel.fit(X_train, y_train)
```

Figure 6. Data is split into testing and training. It is then fit into the model defined from the SciKit Learn library.

The final step in the logistic regression fit is to get back the predictions from the actual model and evaluate the networks performance, as shown in Figure 7.

```
In [75]: #get predictions
         predictions = logmodel.predict(X_test)
         predictions
```

Figure 7. Predictions are derived via “model.predict()” function.

Results

As seen in Figures 8, the network performed extremely well and is an excellent representation of the power of machine learning, especially when it comes to biological implementations such as genetics.

```
In [76]: #Get classification Reports
         from sklearn.metrics import classification_report, confusion_matrix
         print("Model Statistics")
         print(classification_report(y_test,predictions))
         print("Confusion Matrix")
         print(confusion_matrix(y_test,predictions))

         Model Statistics
                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00         4
                    1       1.00      1.00      1.00         5

             accuracy                           1.00         9
            macro avg       1.00      1.00      1.00         9
         weighted avg       1.00      1.00      1.00         9

         Confusion Matrix
         [[4 0]
          [0 5]]
```

Figure 8. Classification report reveals 100% accuracy over limited database.

Discussion

While the models were good predictors, these results may be exclusive to this experiment. The relative importance of each gene may differ from other experiments as gene expression varies from each individual. The issue remains that data regarding differential expression of genes in MS has poor overlap across experiments and this could hinder the development of an accurate algorithm for diagnosing the disease. To improve the NN, more genes are to be evaluated by the model in order to gain a greater scope into which genes may be associated with MS. Larger datasets of both healthy and MS patients are needed in order to improve the decision and overall outcome of the NN. A simple and relatively easy to use model was created and used successfully to identify key genes that may be a cause for MS as well as a predictor for MS based on a small sample of genes from PBMC microarray data. As research in the field of the use of machine learning and genomics continues to expand, we expect to see greater improvements to the model that was successful in this study.

Conclusion

While the models were good predictors, these results may be exclusive to this experiment. The relative importance of each gene may differ from other experiments as gene expression varies from each individual. The issue remains that data regarding differential expression of genes in MS has poor overlap across experiments and this could hinder the development of an accurate algorithm for diagnosing the disease. To improve the model, more genes are to be evaluated by the model in order to gain a greater scope into which genes may be associated with MS. Larger datasets of both healthy and MS patients are needed in order to improve the decision and overall outcome of the model. However, the methodology in this experiment from the data accumulation to the actual construction and testing of the model itself serves as strong representation of the value artificial intelligence can have in the field of genetics.

Acknowledgments

I would like to thank Mr. Kuan-Chen Wu for guiding me down the pathway of using artificial intelligence in various applications Without his guidance and support in continuation of this project, the publication of this document would not have been possible.

References

[1] Berer, Kerstin et al. (18/04/2014). Microbial view of central nervous system autoimmunity. FEBS Letters. https://doi.org/10/1016/j.febslet.2014.04.007. Retrieved: 01/16/2022.
[2] Podbielska, Maria. (08/08/2020). Distinctive Sphingolipid patters in chronic multiple sclerosis lesions. Research Gate. https://www.researchgate.net/figure/Plaques-morphology-in-MS-cases-examined-Tissue-sections-were-stained-with-Luxol-fast_fig1_343519143. Retrieved: 01/16/2022.
[3] Ramteke, Rakesh et al. (03/12/2012). Automatic Medical Image Classification and Abnormality Detection Using K Nearest Neighbor. Research Gate. https://www.researchgate.net/publication/305403850. Retrieved: 01/16/2022.
[4] Zhang, Yudong. (09/06/2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection. Sage Journals. https://journals.sagepub.com/doi/abs/10.1177. Retrieved: 01/16/2022
[5] National, Multiple Sclerosis Soc. (05/08/2017). What Causes MS? National Multiple Sclerosis Society. https://www.nationalmssociety.org/What-is-MS/What-Causes-MS Retrieved: 01/16/2022.
[6] Acquaviva, Massimo et al. (21/07/2020). Inferring Multiple Sclerosis Stages from the Blood Transcriptome via Machine Learning. National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/33205062/. Retrieved: 01/16/2022.
[7] Faria, Omar et al. (22/01/2013). MicroRNA dysregulation in multiple sclerosis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3551282/. Retrieved: 01/16/2022.
[8] Nehal M, Ali et al. (10/10/2020). Machine Learning In Early Genetic Detection of Multiple Sclerosis Disease: A Survey. International Journal of Computer Science and Information Technology. https://aircconline.com/ijcsit/V12N5/12520ijcsit01.pdf. Retrieved: 01/17/2022
[9] Institution, Genome. (18/01/2017). Microarray Technology. National Human Genome Research Institution. https://www.genome.gov/genetics-glossary/Microarray-Technology. Retrieved: 01/17/2022
[9] Institution, Genome. (18/01/2017). Microarray Technology. National Human Genome Research Institution. https://www.genome.gov/genetics-glossary/Microarray-Technology. Retrieved: 01/17/2022
[10] Moore, Craig S. (22/01/2013). MicroRNA dysregulation in multiple sclerosis. Frontiers In Genetics. https://doi.org.10.3389/fgene.2012.00311. Retrieved: 01/16/2022.