Neural Nets and Blood Samples: A Novel Way of Identifying Drug Resistant Epilepsy Patients

Tony Xu[1], Liu Jinpeng[2]

[1]Math, Science, and Technology Center
Paul Laurence Dunbar High School
Lexington, KY 40513

[2]College of Medicine
University of Kentucky
Lexington, KY 40526

May 8, 2024

Abstract

One in twenty-six people worldwide develops epilepsy during their lifetime. Around 30% of epilepsy patients are nonresponsive to medication. It is important to identify these patients using less invasive and low-cost methods. This study aims to use gene expression profiles derived from blood cells and previously identified differentially expressed genes to create a predictive neural network model that can differentiate between drug responders and non-responders.

To determine whether expression profile changes in blood samples can reflect epilepsy pathogenesis, patients with epilepsy were grouped based on their response to medication. Responders and non-responders were compared with each other, establishing a list of differentially expressed genes. A subset of genes within that list was used as inputs to construct neural networks which were used to predict the drug response, to find a network with an accuracy of at least 90%.

Results showed that from the list of differentially expressed genes, decreasing the number of inputs produced a noticeable yet statistically insignificant difference in the accuracy of the model. Differences in the number of samples used to train the model did significantly decrease accuracy with smaller sample sizes. Some models did produce models with an average of 90% accuracy, providing the basis for the usage of gene signatures in peripheral blood samples to predict drug response in epilepsy patients.

Table of Contents

Introduction

Epilepsy is a disease that is defined by recurring seizures, affecting tens of millions of people across the world, including myself. Many with epilepsy seek treatment in the form of drugs, which may or may not be effective in preventing seizures. One in three epilepsy patients have seizures that are untreatable by drugs. Once a patient has taken two medications, there is only a 5% chance a third medication will be effective [1].

Therefore, it is important to identify biomarkers to predict drug response in epilepsy patients. Furthermore, such biomarkers must be obtained as non-invasively as possible with as little cost and time needed to process as possible. Of all these methods, peripheral blood samples taken from epilepsy patients after treatment were identified as one possible source of data for biomarker information.

Within blood samples, different mRNAs can be found that are associated with the expression of different genes. These biomarkers can be compared between different phenotypes using differential expression analysis, which uses a t-test to compare the difference in expression between two groups of samples and identify statistically significant differences in expression. Differential expression also identifies the direction in which a certain gene is differentially expressed and whether it is upregulated or downregulated in a certain population compared to another. Simultaneously, multiple methods of identifying the role of groups of genes within the body have also been developed, notably pathway enrichment, and gene ontology or GO enrichment analysis, which can identify enriched signaling pathways and gene ontologies that are overrepresented in a dataset. This can help identify the significance of differentially expressed genes within the body.

The purpose of this study is to identify differentially expressed genes between drug-resistant and non-resistant patients for multiple drugs using gene expression profiles derived from blood samples of epilepsy patients. Furthermore, this study also seeks to understand the nature of these differentially expressed genes within the body and how they work with other genes in signaling pathways. Furthermore, the goal of this study is also to identify the possibility of creating a predictive model that can determine if a patient is responding to medication or not just using data gathered from mRNA sequencing. The ultimate aim of this study is to determine if peripheral blood samples are viable data sources to predict drug response in epilepsy patients.

Methods and Materials

Data Preprocessing

The data was taken from the Gene Expression Omnibus, an online database of gene expression profiles (Rawat 2020). The data contained 141 deidentified samples, including healthy controls (n = 50), drug-naive patients (n = 34), drug responders (n = 33), and drug non-responders (n = 24). This data was uploaded and analyzed in RStudio, an IDE for the R language.

The probes in the expression analysis were first matched with gene names, with duplicate probes removed. The data was then normalized using a logarithmic expression with base 2, and replicate samples were then removed from the samples as well. The data was divided into two datasets, one containing untreated individuals (drug-naive and healthy control samples), and one containing individuals treated with medication (drug responders and non-responders).

Differential Expression Analysis

Data from the dataset containing treated individuals was inputted into a differential expression analysis using the limma package (Ritchie 2015). Genes with a p.value less than 0.05 for differential expression were separated into a separate gene list and categorized as differentially expressed. The dataset with treated individuals was then trimmed, removing any gene that was not differentially expressed. The resulting dataset was used to train and test the neural networks.

Training the Neural Network

Formatting Data

The trimmed dataset was split into two dataframes with two sets of samples, a larger one containing 47 samples used to train the model and a smaller testing dataset used to gauge the accuracy of the model containing a subset of 5 drug responders and 5 drug non-responders. These dataframes were transposed so that the samples became rows instead of columns. A phenotype column was added in each dataframe as either “Responder” or “Non-responder” to let the neural network distinguish between the two in the samples.

Selecting Parameters for the Neural Network

The neural network was trained using the neuralnet package (Fritsch 2022). Large multivariable neural networks were unsupported by the package, so the number of genes used in each neural network was limited. The parameter predicted was the phenotype column for each sample. The number of genes to be used as inputs ranged from 60 genes to 180 genes in intervals of 30 (60, 90, 120, 150, and 180 genes) for 5 groups of neural networks. Two hidden layers were included in each model, the first containing 20 nodes, and the second containing 5 nodes. Two nodes, “Responder” and “Non-responder” were contained in the output.

Once the parameters were selected, each neural network would train on the training dataset, finishing once the accuracy of the error was minimized, returning an object containing the weights for each node in the neural network.

Testing the Neural Network

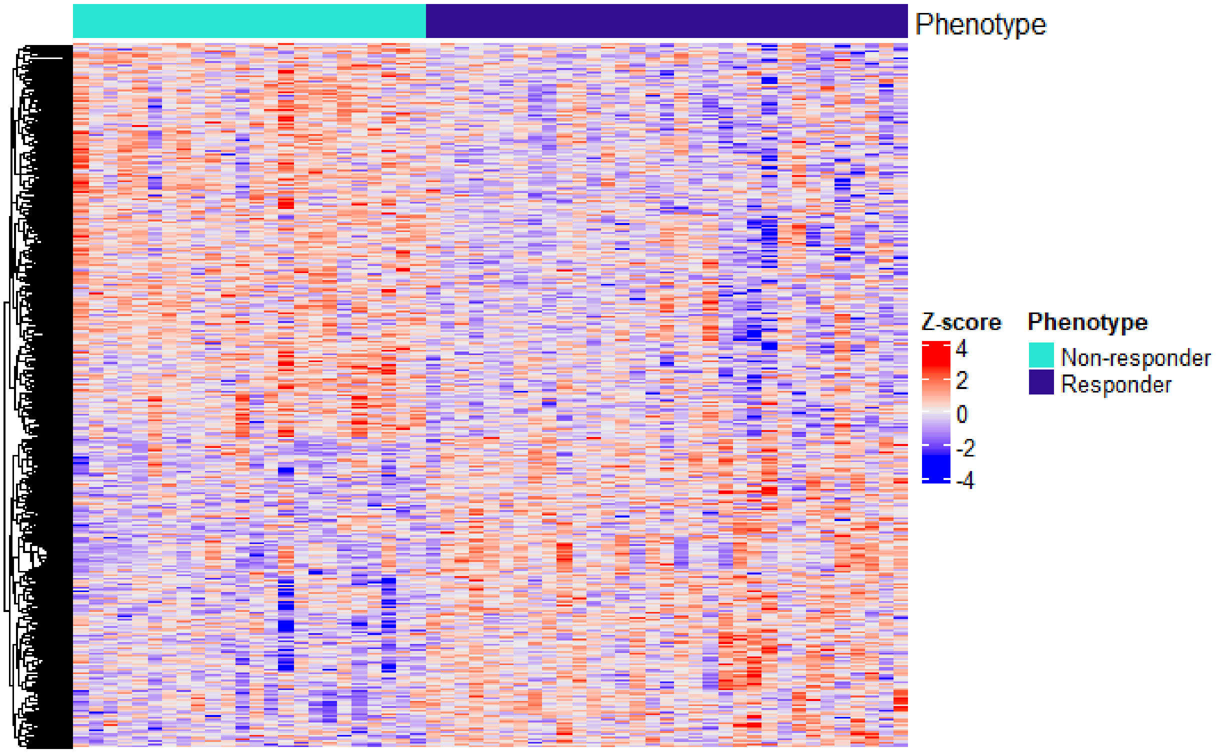
5 neural networks were created for each group of neural networks, for a total of 25 neural networks. The samples in the testing dataset were used to test the neural network, and a confusion matrix generated using the caret package (Kuhn 2023) returned the results in a confusion matrix. The accuracy was returned as a numerical value between 0 and 1 in the confusion matrix, which was averaged across all 5 neural networks for each group of neural networks.

Results

Differential expression analysis identified 1104 genes as differentially expressed between responders and nonresponders (Figure 1) using a p.value cutoff of 0.05 for statistically significant differential analysis.

Figure 1

Heatmap of Differentially Expressed Genes Identified with Differential Expression Analysis



Note. Blue areas indicate underexpression and red areas indicate overexpression. A clear difference between responders and nonresponders can be seen. A subset of these genes were used as inputs to train the model.

The confusion matrices generated by the predictions of the neural network produced accuracy values for each individual neural network in a group. Each group contained 5 neural networks over 5 trials. The results of the neural network predictions (Table 1) demonstrated that as the number of gene inputs increased, there was a statistically insignificant decrease in accuracy from 0.9 to 0.82 (ANOVA test, F = 0.24468) with large variations in the accuracy between trials in each group (Figure 2).

Table 1

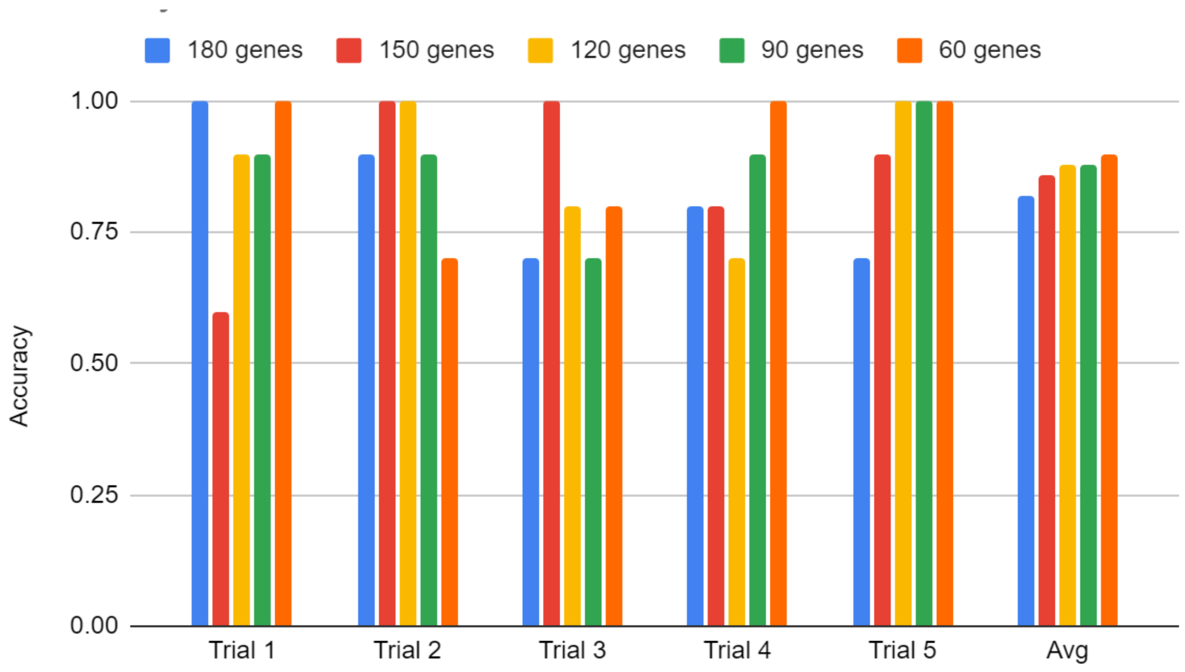
Accuracy of Neural Networks over 5 Trials with Different Numbers of Gene Inputs

Trial	Number of Gene Inputs				
	60	90	120	150	180
1	1	0.9	0.9	0.6	1
2	0.7	0.9	1	1	0.9
3	0.8	0.7	0.8	1	0.7
4	1	0.9	0.7	0.8	0.8
5	1	1	1	0.9	0.7
Average	0.9	0.88	0.88	0.86	0.82

Note. According to an ANOVA test, the differences are not statistically significant (F = 0.24468)

Figure 2

Graph of Neural Network Accuracy over 5 Trials with Different Numbers of Gene Inputs

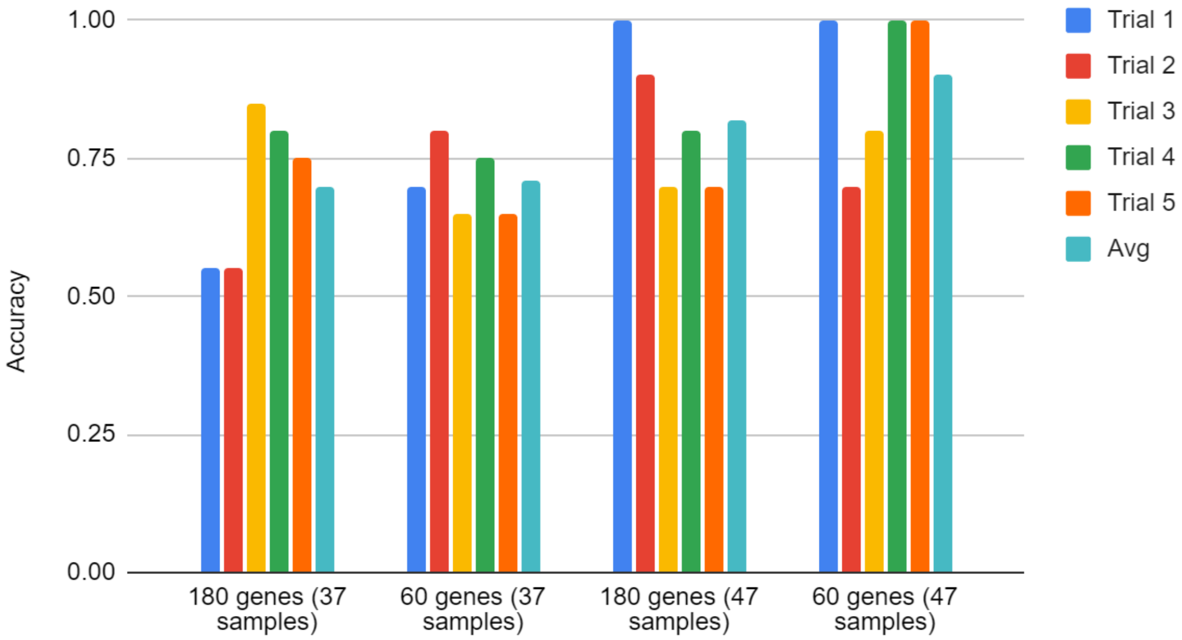


Note. A clear upwards trend can be seen in the averages columns as the number of genes in the network decreases, though by an ANOVA test, this is not statistically significant (F = 0.24468)

Confusion matrices accuracy values also showed that when the number of samples was decreased in the 60 input and 180 input groups, the accuracy of the model decreased to 0.71 and 0.7, a difference that was statistically significant in the 60 inputs groups (two sample t-test, p = 0.0259) but not in the 180 input groups (p = 0.12) (Fig. 3). When compared together, by unifying the 60 input and 180 groups into a single group with 100 neural networks, the difference in accuracy was statistically significant (p = 0.01).

Figure 3

Graph of Neural Network Accuracy over 5 Trials with Different Numbers of Gene Inputs and Different Numbers of Training Samples



Note. The differences between the 180 gene groups were insignificant (p = 0.12) but the difference between the 60 gene groups and between a combined group with both 180 and 60 genes were statistically significant (p = 0.0259 and p = 0.01 respectively)

Discussion

The data in this experiment provided partial support for the hypothesis. The trends found in predictive accuracy by varying the number of inputs ran counter to the hypothesis. However, more trials are needed to establish any significant difference between the groups of neural networks. The accuracy results from decreasing the size of the training dataset not only supports the hypothesis, it also indicates that like many neural networks, the robustness and accuracy of neural networks are significantly increased with the addition of more samples.

The data in this study provides insight into the possibility of using blood samples to predict drug response in epilepsy patients. Though the networks created were still not viable for use in a clinical setting, the accuracy with a small dataset is still promising. Peripheral blood samples have also been established as a new area for the study of epilepsy drug response with the discovery of many genes that were differentially expressed between responders and non-responders.

One of the biggest limitations this study faces is a lack of data in the training data set. In order to prevent overfitting of the data, the samples were removed from the already small original dataset to make a testing dataset. Even in demonstrations of the neuralnet package used to generate the neural networks, the sample datasets contain at least 100 samples in the training set, and the 47 samples used to train most of the neural networks is likely not enough to make a robust model. A dataset with hundreds or even thousands of samples could be used to train more robust neural networks with higher predictive accuracy.

Homogeneity is also a limitation of this study. The dataset contained patients who were put on 3 different treatments: Carbamazepine, Valproate, and Phenytoin. The mechanisms for these drugs are different, however separating them would have resulted in a training dataset that was too small to draw meaningful results from. A larger dataset to train the neural networks has to also be homogenous to eliminate any confounding variables in analysis.

Current literature on the genetics of drug resistance mostly focuses on gene expression in the brain tissue of patients [1][2]. However, a genetic study in 2013 found that there were differentially expressed mRNAs in the blood samples of epilepsy patients and healthy controls without epilepsy [3]. This study extends the study of mRNA expression patterns and blood samples as well as the study of the genetics of drug resistant epilepsy by looking at the differentially expressed genes of epilepsy patients who respond and those who do not respond. In the future, these networks will also be trained with more and more complex methods, like deep learning and other advanced machine learning algorithms. More epilepsy patients' blood samples will be tested for mRNA expression, as these algorithms will require larger datasets and will also likely need to be more homogenous. More datasets can also be made for individual drug treatments and also for different time-points in treatments. For example, blood samples can be taken before a patient is put on medication in order to identify if specific biomarkers are differentially expressed between responders and nonresponders before they even take medication.

Acknowledgements

This research was made possible by my friends and family who have supported me throughout this research journey. I would also like to thank the MSTC Program and Mrs. Elizabeth Pelphrey for guiding me in my research and Dr. Liu Jinpeng for mentoring me and helping me understand my research better.

References

1. “Drug-Resistant Epilepsy Facts.” NeuroPace, Inc., 2019, https://neuropace.com/patients/epilepsy-infographics/drug-resistant-epilepsy-facts/.
2. Alberto Lazarowski, Gustavo Sevlever, Analía Taratuto, Mario Massaro, Adrián Rabinowicz, Tuberous sclerosis associated with MDR1 gene expression and drug-resistant epilepsy, Pediatric Neurology, Volume 21, Issue 4, 1999, Pages 731-734, ISSN 0887-8994, https://doi.org/10.1016/S0887-8994(99)00074-0.
3. Xi, Z. Q., Xiao, F., Yuan, J., Wang, X. F., Wang, L., Quan, F. Y., & Liu, G. W. (2009). Gene expression analysis on anterior temporal neocortex of patients with intractable epilepsy. Synapse (New York, N.Y.), 63(11), 1017–1028. https://doi.org/10.1002/syn.20681
4. Greiner, H. M., Horn, P. S., Holland, K., Collins, J., Hershey, A. D., & Glauser, T. A. (2013). mRNA blood expression patterns in new-onset idiopathic pediatric epilepsy. Epilepsia, 54(2), 272–279. https://doi.org/10.1111/epi.12016