Predicting Shared Risk Genes between Autism, CHD, and Schizophrenia using Gradient-boosted Decision Trees

By: Katrina Zhao, Evelin Leiva, Tanmay Jaiswal COMS4762, April 24, 2021

Abstract

While sequencing is becoming easier and more affordable over time, gene expression data from specific regions of tissues are still relatively difficult to come by, resulting in a lack of data for certain disorders or diseases people may want to study, this is especially the case for de novo associated gene mutations, which are vastly not yet understood. We propose a method of using neural networks with feature extractors to combine data from a variety of sources, even if in some cases, the number of features, the length of the data, or the sources do not align completely. Here, we use a model that feeds the outputs from a feature extractor neural network into a gradient boosted tree on combined scRNA-seq data from a variety of sources to find common risk genes between congenital heart disease, autism and schizophrenia. Our model was able to successfully classify risk genes for autism and schizophrenia, but was unable to make significantly accurate predictions for congenital heart disease. We found that, at least in our datasets, there are twelve risk genes in common between all of our disorders at 50% confidence and three risk genes in common between all of our disorders at 70% confidence.

Introduction

With the success of machine learning models in various projects from detecting autism spectrum disorder to predicting cardiovascular disease, our project is based on exploring the shared and distinct risk genes of three disorders known to be associated with de novo mutations, many of which have yet to be fully mapped or understood. Congenital heart disease (CHD) and neurodevelopmental disorders like autism have been shown to be correlated with the expression of similar de novo mutations during embryonic-stages of development, while some mental disorders such as schizophrenia, which typically develops during adulthood, have similarly been linked with sets of de novo mutated genes as well. Thus, we want to investigate their relationship, like whether this association holds between the three disorders together and if it is through similar sets of genes.

In their paper “Dissecting Autism Genetic Risk Using Single-Cell RNA-Seq Data,” Chen et. al., use a method called A-risk, which is based on gradient boosted decision trees, to investigate which genes are associated with a higher risk for autism. The A-risk method learns known autism risk genes’ expression patterns and then predicts the likelihood that any gene is an autism risk gene. We would like to adapt this method to predict risk genes in three disorders - autism, CHD, and schizophrenia - to

find any common risk genes between them. To do so, for each disorder independently, we trained our data with a custom feature extractor neural network, and used our outputs from that as inputs for our boosted tree based on the A-risk boosted tree model. Then, we manually compared these predictions for the risk genes associated with each disorder to find if there was any overlap.

Methods

Our methodology was predominantly inspired by the A-risk method described in the Chen et al. paper and from programming assignment code in this class. This translated to creating a model able to train gradient-boosted trees with each disorder- autism, CHD, and schizophrenia - independently, despite the data stemming from different projects. To begin, we extracted scRNA-seq data from developing human brain tissue as one set of inputs and pre-processed the corresponding autism risk genes from Chen et al's paper. For CHD, we extracted scRNA-seq data from developing human heart tissue from the paper "Congenital heart disease risk loci identified by genome-wide association study in European patients" (Lahm et. al, 2021) and processed CHD risk genes from the paper "Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands." (Jin, Sheng Chih, et al. 2017). Finally, because the onset of schizophrenia symptoms and subsequent diagnosis largely occurs during adulthood instead of during early development, we also obtained scRNA-seq data from adult human brain tissue from the paper "Multi-scale analysis of schizophrenia risk genes, brain structure, and clinical symptoms reveals integrative clues for subtyping schizophrenia patients" (Liang Ma et al. 2019) and pre-processed schizophrenia risk genes from "De Novo Mutations Identified by Exome Sequencing Implicate Rare Missense Variants in SLC6A1 in Schizophrenia." (Rees, Elliott, et al. 2020).

For our model, we used a two-step architecture. The first part of our architecture consists of a neural network feature extractor. The second part involves feeding the output from the feature extractor to a gradient boosted tree.

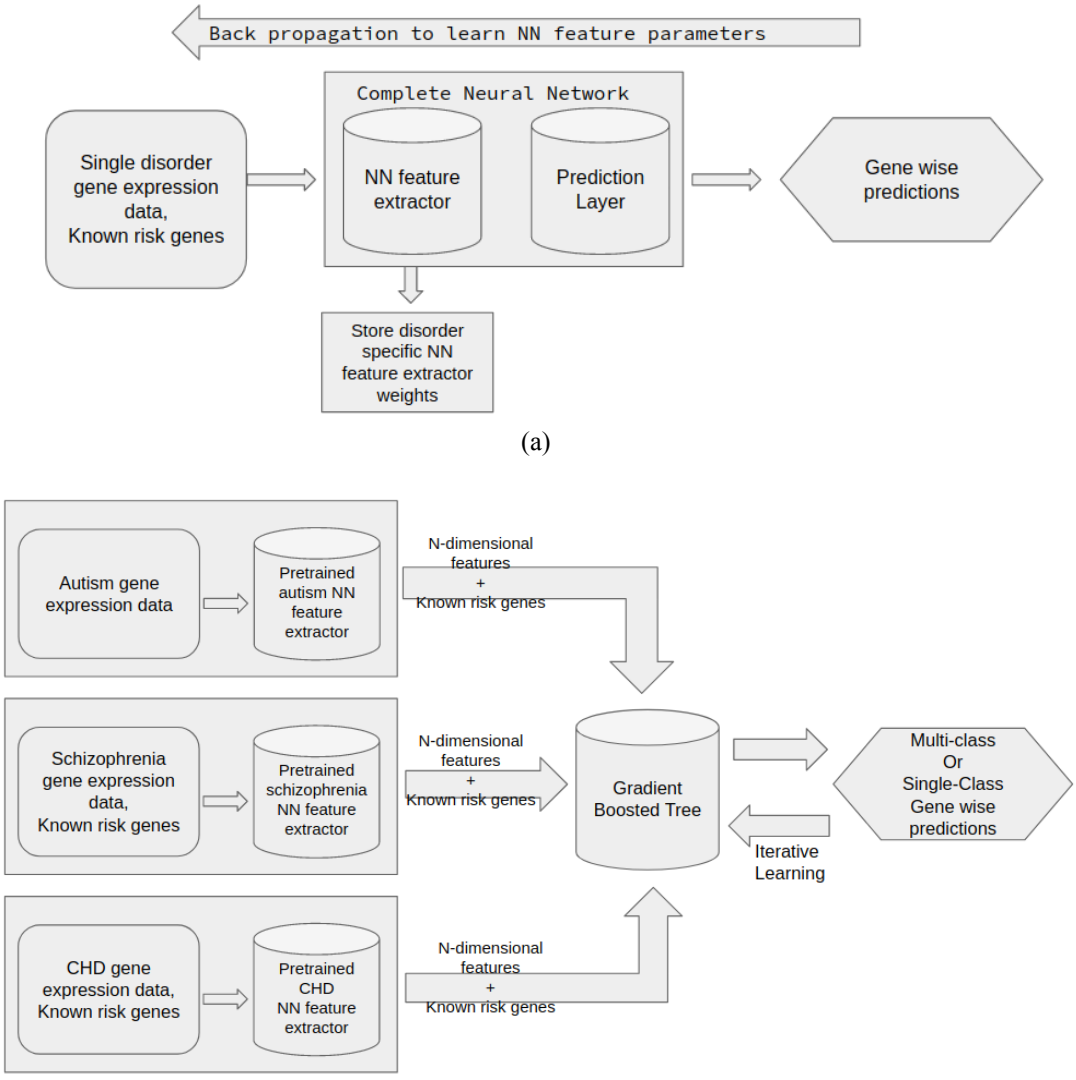

Figure 1: Solution architecture
(a) During the Pre Training phase each feature extractor is trained as part of a simple NN
(b) During the Training phase the feature extractor creates N-dimensional features for the Gradient Boosted Trees

The neural network is a simple two layer model with the input size equal to the number of parameters in the data from the disorder analyzed. The output layers across the three disorders are of consistent size so that the inputs to the gradient boosted tree have the same number of parameters. Our RNA-seq data was taken from different cell types within the same tissue, so our data has high dimensionality. The purpose of the neural network feature extractor is to create non-linear functions which are combinations of the various input parameters. Thus, we can model relationships between the various cell types within each RNA-seq dataset with the neural net. Additionally, the neural network uses ReLU activation, dropout, and batch normalization. We found that adding both dropout

and batch norm results in a marked improvement in the performance. Dropout forces the feature extractor to learn new features by removing some data that it may heavily rely upon. The batch normalization makes the performance of the gradient boosted tree independent of the scale of the outputs from the neural network since the outputs are normalized.

A neural network was trained for each of the three disorders. Each neural network was coupled with a simple output layer and trained independently in a supervised setting. The best performing neural network model was chosen from a 5-way cross-validated dataset after training five times with a new random initialization of weights. The problem was phrased as a simple classification task, and cross entropy loss was used. We used class weighted cross entropy loss to compensate for the class imbalance in our data. Thus, the loss was weighted as the inverse of the frequency of occurrence of a class. This resulted in higher precision and recall scores instead of simply prioritizing accuracy since accuracy is a misleading metric with this level of class imbalance.

Finally, we trained the gradient boosted tree classifier based on A-risk with the outputs from our best performing neural network model. The gradient boosted tree had three hundred boosted stages and only one level of depth. We found that further depth resulted in increased overfitting and hence lower validation losses. Finally, in order to determine which risk genes were in common between the three disorders based on our predictions, we compared the three lists of disorder positive genes generated by our model to find overlapping ones.

Results

After training our neural network on our autism RNA-seq data, we generated accuracy curves for the training and validation sets.

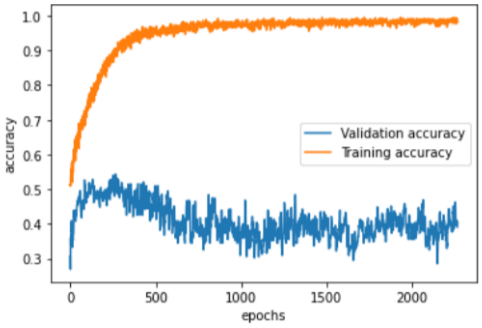

Figure 2: Accuracy curve for autism risk genes from feature-extractor NN

We found that the validation accuracy was low (0.58) when we train with a hidden size of 60. However, we wanted to choose a hidden size that was common across the models for all of our disorders in order to eventually be able to multitask with our gradient boosted tree in the future. A hidden size of 60 was optimal for the schizophrenia data, but caused our autism neural network to overfit. From the

curves, we can see that the accuracy eventually decreases, so we selected the model that gave us the highest accuracy in an earlier epoch to use with our gradient boosted tree. After running the gradient boosted tree, we found that our model actually performed very well. From the 5-fold cross-validation ROC curve, we can see that there's a high rate of true positives across false positives rates in our prediction. In fact, in comparison to the predictions from the A-risk paper, we actually had a higher mean AUC (0.91 compared to their 0.77). Most likely this is because we used a feature extractor in addition to the gradient boosted tree. Thus we could model relationships between the different cell types within the brain tissue and account for any effect they have on gene expression. The original A-risk paper misses out on this complexity of the interplay that decides the macro-level impact of the expression of the genes.

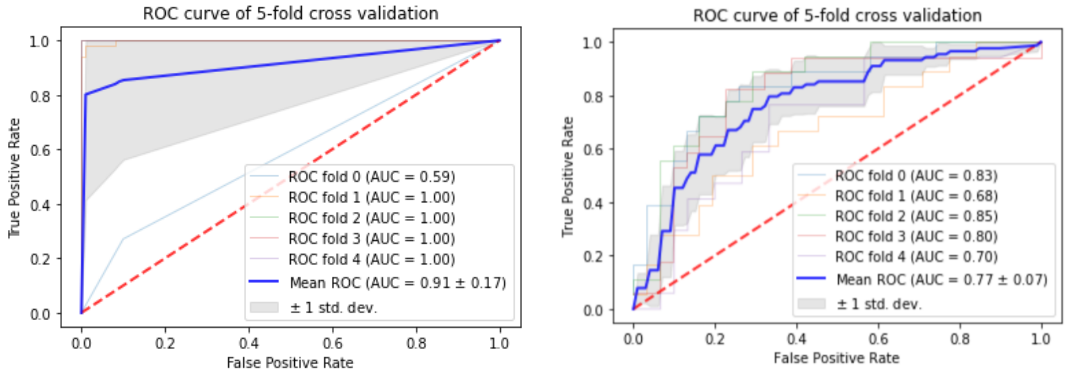

Figure 3: Autism ROC curve feature-extractor NN + GB tree model (left) and A-risk model (right)

We then attempted to train our model on the data for CHD. However, after generating the accuracy plots of the neural network in the first part of our model and then examining the loss model subsequently, we found that the performance was severely lacking.

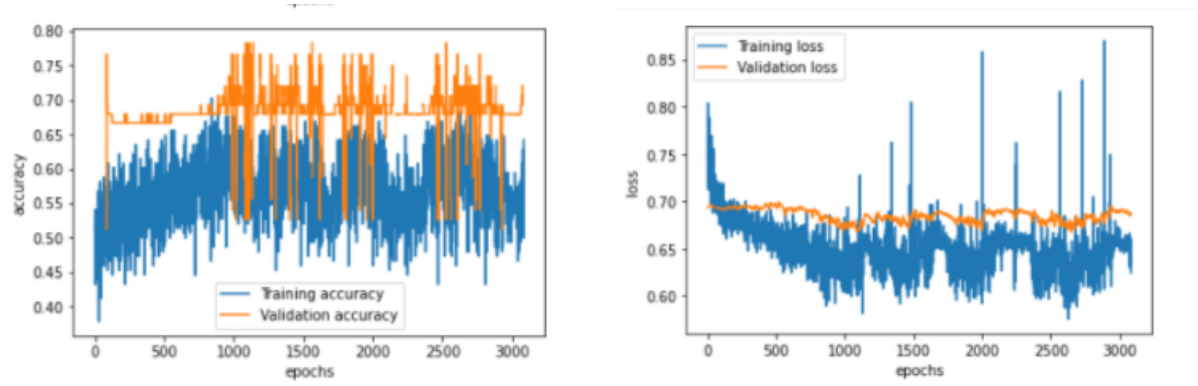

Figure 4: Accuracy (right) and loss curves (left) for CHD risk genes from feature-extractor NN

The loss never seems to converge even after many epochs, and the accuracy curve is a plateau. These results are most likely due to the data imbalance mentioned earlier between our CHD class earlier.

The ratio of CHD positive genes to CHD negative genes is very low. We attempted to improve our predictions by using weighted cross entropy loss, oversampling the minority class, and strome oversampling. However, none of these methods improved our results significantly. After training the gradient boosted tree part of our model, we see that indeed our predictions are not much better than random prediction as our mean AUC was 0.51.

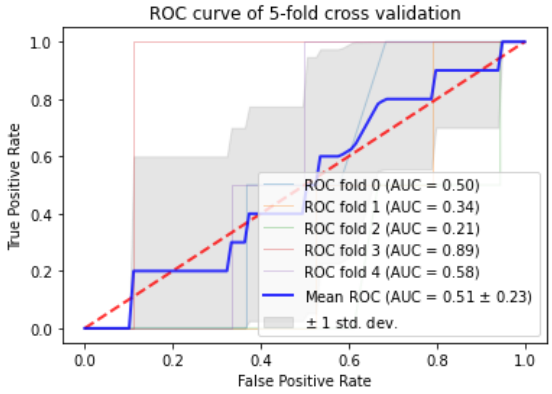

Figure 5: CHD ROC curve with feature-extractor NN + GB tree model

Finally, we trained our model on schizophrenia data. After generating the accuracy curve, we found our model had a maximum accuracy of 0.72 and stays at a high accuracy after many epochs.

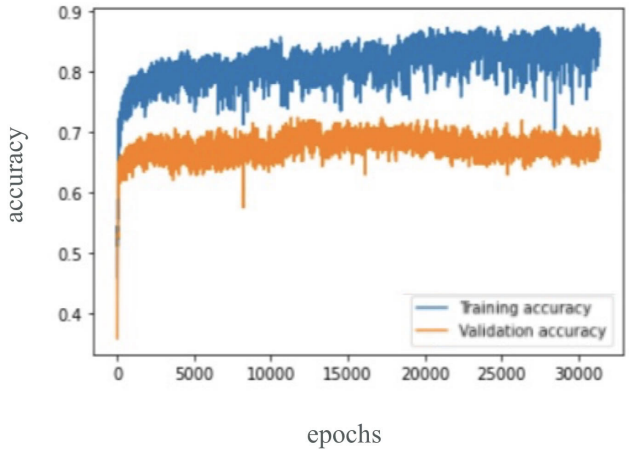

Figure 6: Accuracy curve for schizophrenia risk genes from feature-extractor NN

After training with the gradient boosted tree, we see that similar to autism, our model also seems to predict with a high rate of true positives over false positives rates. From the ROC curve of the 5-cross validation, our mean AUC was 0.88.

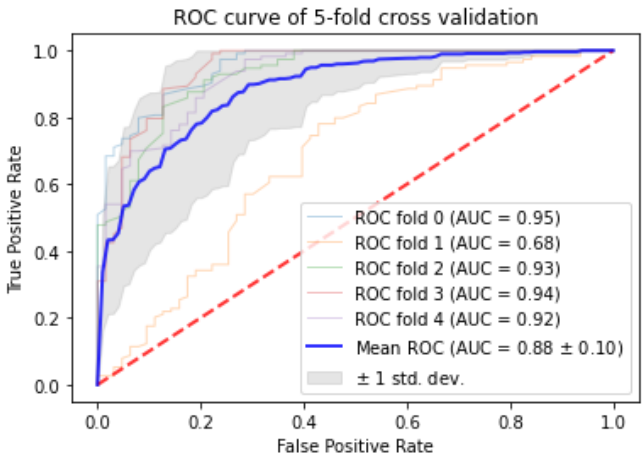

Figure 7: Schizophrenia ROC curve with feature-extractor NN + GB tree model

After making all of our classifications, we looked at the sets of genes that were positive for each disorder and tried to find the genes that were in common. We found a total of twelve risk genes across all three of our disorders (GRIP1, ACVR2B, CTNND2, NMNAT2, DSCAM, CACNA2D3, SHANK2, CNTNAP2, NLGN3, KATNAL2, CTNNA2, and RELN) at 50% confidence. There were three risk genes across all three disorders (GRIP1, KATNAL2, and ACVR2B) at 70% confidence. Interestingly, in our initial dataset from our cited papers, GRIP1 and KATNAL2 were not linked to CHD, and ACVR2B was not linked to autism. However, other literature indicates that these three genes are linked to each of the disorders. Additionally between autism and CHD only there were 26 common risk genes at 90% confidence. Between autism and schizophrenia only there were 41 common risk genes at 90% confidence, and between CHD and schizophrenia only there were 8 common risk genes at 90% confidence. The names of these genes are given in the appendix. Of course, our results from the CHD datasets must be considered with caution due to poor performance of our model on the little data that we did have. However, looking into the pathways that these common risk genes are involved in could provide more insight on the similarities between the pathogenesis of these three disorders.

Our code can be found here: Google colab repo Link

Discussion

Our initial expectation was that there would be a considerable group of genes common to CHD, autism, and schizophrenia that we could link, but our results lead us to the conclusion that there are actually only 3 genes that we can determine at 70% confidence with our current datasets. This leads us to believe that perhaps we should investigate a wider set of genes found in each to see if there are other common unintuitive genes and expand our data pool. Our project did lead to a high accuracy in determining risk genes for two out of three disorders with the data that we did have however.

Since there has not yet been a prior study that compared these three disorders with each other, we had a challenge in finding data that gave gene expressions in humans and corresponding risk genes for each disorder, and at corresponding stages of life. Particularly, obtaining formatted and publicly accessible scRNA-seq from embryonic human heart data and corresponding risk genes that would lead us to sizable amounts of genes for congenital heart disease after processing and filtering proved to be a challenge, which could be tied to the input data focusing on subjects of only European descent. While we were able to implement a new method for determining risk genes in each of our target disorders and were successful in improving performance on determining autism and schizophrenia genes, it was difficult to work with our data as it stemmed from different sources with differing methods, features, and subject focuses. We believe that access to more expansive data could lead our model to improve its performance. To take this work even further, we would like to create a method to train our gradient boosted tree to develop indexes for all three disorders in parallel in order to multitask-version of our existing model, which was our original aim of our project proposal. Given more time, we would be able to accomplish this since we have already built the architecture for the multitasking gradient boosting tree and just need to work through some minor errors from piping in the data from the three neural networks to the tree. We would then analyze the subsequent results for differences between the independent models and the multitask model, comparing performance.

Nonetheless, our project is able to successfully create an implementation that can unite gene expressions and risk gene data from different sources via a custom embedding that improves overall accuracy. Our adaptation of the original A-risk method is able to run on congenital heart disease data and schizophrenia, which has not yet been done before. Given more access to training data, our developed method can prove useful in further investigation of de-novo mutation associated disorders. Our results also can inspire a subsequent investigation of the importance of features on each disorder to further study which are most important for each disorder. In either sense, our project succeeds in getting one step closer to understanding the role of de-novo mutation genes in these disorders.

References

Chen, Siying, et al. “Dissecting Autism Genetic Risk Using Single-Cell RNA-Seq Data.” *Cold Spring Harbor Laboratory*, 16 June 2020, p. 2020.06.15.153031, doi:10.1101/2020.06.15.153031.

Jin, Sheng Chih, et al. “Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands.” *Nature Genetics*, vol. 49, no. 11, Nature Publishing Group, Oct. 2017, pp. 1593–601.

Kessler, Michael D., et al. “De Novo Mutations across 1,465 Diverse Genomes Reveal Mutational Insights and Reductions in the Amish Founder Population.” *PNAS*, National Academy of Sciences, 4 Feb. 2020, www.pnas.org/content/117/5/2560/tab-article-info.

Lahm et. al, “Congenital heart disease risk loci identified by genome-wide association study in European patients” *Journal of Clinical Investigation” 2021;131(2):e141837. https://doi.org/10.1172/JCI141837.*

Liang Ma et al., “Multi-scale analysis of schizophrenia risk genes, brain structure, and clinical symptoms reveals integrative clues for subtyping schizophrenia patients”, *Journal of Molecular Cell Biology*, Volume 11, Issue 8, August 2019, pp 678–687,

Rees, Elliott, et al. “De Novo Mutations Identified by Exome Sequencing Implicate Rare Missense Variants in SLC6A1 in Schizophrenia.” *Nature Neuroscience*, vol. 23, no. 2, Nature Publishing Group, Jan. 2020, pp. 179–84.

Yuan, Ye, and Ziv Bar-Joseph. “Deep Learning for Inferring Gene Relationships from Single-Cell Expression Data.” *PNAS*, National Academy of Sciences, 26 Dec. 2019, www.pnas.org/content/116/52/27151/tab-article-info#corresp-1.

Appendix of Common Genes between Disorder Pairs at 90% Confidence:

Autism and Schizophrenia:

GRIA1, GRIP1, MECP2, SRSF11, CACNA1H, DIP2C, NRXN1, ASXL3, SHANK2, RANBP17, DNMT3A, MED13, KMT2C, SMARCC2, CNTN4, CACNA2D3, PHF3, KATNAL2, CHD2, RELN, INTS6, CUX1, SET, SHANK3, TAOK2, TCF20, FOXP1, UBN2, KAT2B, CIC, ANKRD11, WAC, GRIN2B, BCOR, CEP57, SETD5, SUV420H1, RASA1, GIGYF2, KDM6A, POGZ

Autism and Congenital Heart Disease:

SPDL1, SLC4A1, RGS5, POLE, ADCY1, SMOC1, KIF26A, NIPSNAP1, CCBE1, GCK, PRR7, AMPH, MYO1B, PRKCB, GYLTL1B, PABPC1, CTNNA2, GEM, WNK3, CCP110, SLC6A4, PRDM12, CELSR3, CHST10, ARNT2, FERMT3

Schizophrenia and Congenital Heart Disease:

SOX2, JAG1, TMEM67, SEMA3E, PLXND1, NOTCH1, ACTB, PHGDH