REPORT

Exploring Brain-Neurological Problems Links with Multi-Omics and Explainable AI

Vivekanand R (Sample Report for Github Viewers on Gene Classification Tasks - Only for Viewing)

Submitted on November 7, 2024
for the Unknown Course
in the Error

Abstract Many large-scale initiatives and health associations focus on disorders like Alzheimer's, Parkinson's, and schizophrenia, leading to publicly available multi-omics data sets. In this report, we will integrate and explore multi-omics data (including genomics) for predicting complex disease traits. Also, utilizing Explainable AI models to derive actionable connection between human gene and its influence on neurological problems. Using explainable AI to highlight the key genes, pathways, or molecular mechanisms that differentiate the potential vulnerable gene.

Keywords Neurological, Brain, Explainable AI, Gene

1 Introduction

In recent years, AI has made a significant progress in many healthcare areas whether its early cancer detection, CT, MRI and even advanced clinical diagnosis to assist doctors on complex surgeries. But still there were few challenges remain unresolved. Today, there is no cure for certain medical healthcare conditions (i.e., Mental Health problems like severe depression, Schizophrenia, Alzheimer and other neurological conditions). Even though, these conditions can be treated, but still there is no permanent cure for these traits. As per World Health Organization (WHO), 1 in 8 people live with some form of mental health condition and in which 71 percent of them remain untreated.

Let's delve into each category:

Alzheimer: Affects aged people (65+), prevalently found around 1 in 10 older people. Some of the risk factors were Memory Decline, Sleep trouble, Genetics, clinical depression, head injuries, hypertension and other psychological stress.

Severe Depression Bi-Polar Disorder: Another major neurological problem. Sudden changes in a person's mood, manic depression could be a symptoms.

Schizophrenia: An another neurological disorder which affects 1 in 100 people (less than 1 percent of world population), Age Range: Mid of twenties until mid thirty's, Some symptoms were out of touch with reality, difficult in concentration, hallucinations, see/hear things that doesn't exist.

2 Overview

As noted in the introduction, the process involves below steps:

2.1 Data Sources and Collection
2.2 Pre-processing Steps
2.3 Methodology
2.4 Explainablity using AI
2.5 Validation and Output

2.1 Data Sources and Collection

Data used in this project were collected mainly from GWAS Catalog (Genome wide associations study - https://www.ebi.ac.uk/gwas/). Three main neurological categories has been identified and focused here.

Below flow chart explains that initially we collected all human gene list (19500+) and 438 different varieties of published research studies been consolidated on these three specific health conditions. And, using explainable AI TSNE method, we have separated gene risk category which has high associations among published studies .

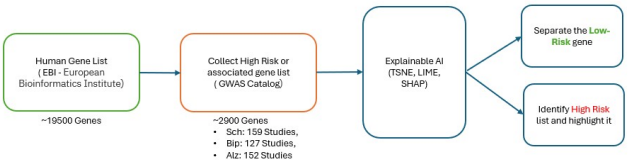

Fig. 1: Flow Chart

Table 1: Frequency of Diseases And its Respective Age Group Estimates

Disorder Category	Frequency Estimates	Age Group
Alzeimer	1 in 10 older people	Affects aged people (65+)
Bipolar	2 to 4 percent of adults	Mostly Adults
Schizopherenia	1 in 100 people	Age mid of 20s and 30s

2.2 Pre-processing Steps

On three categories, gene information and associations were collected from multiple databases. High-risk gene identified on each of three categories (Each data point has been published in Pubmed and other notable publications)

Below are the list of pre-processing Steps:

Missing Gene Values: Data Clean-up been performed.

Normalization and Quality Control: RAF (Risk Allele Frequency) and P-value.

Feature Selection: Dimensionality reduction techniques like PCA, t-SNE been employed and biologically driven feature selection methods been considered to maintain interpretability of this Explainable AI models.

	symbol	name	Risk	Category	Freq	Sch	Alz	bip
0	A1BG	alpha-1-B glycoprotein	No	NaN	0	0	0	0
1	A1CF	APOBEC1 complementation factor	No	NaN	0	0	0	0
2	A2M	alpha-2-macroglobulin	No	NaN	0	0	0	0
3	A2ML1	alpha-2-macroglobulin like 1	Yes	Only Schizophrenia	1	1	0	0
4	A3GALT2	alpha 1,3-galactosyltransferase 2	No	NaN	0	0	0	0
5	A4GALT	alpha 1,4-galactosyltransferase (P1PK blood gr...	No	NaN	0	0	0	0
6	A4GNT	alpha-1,4-N-acetylglucosaminyltransferase	No	NaN	0	0	0	0
7	AAAS	aladin WD repeat nucleoporin	No	NaN	0	0	0	0
8	AACS	acetoacetyl-CoA synthetase	No	NaN	0	0	0	0
9	AADAC	arylacetamide deacetylase	No	NaN	0	0	0	0

Fig. 2: Gene list with risk category - Top 10 Sample View

	symbol	name	Risk	Category	Freq	Sch	Alz	bip
19252	ZYG11A	zyg-11 family member A, cell cycle regulator	No	NaN	0	0	0	0
19253	ZYG11B	zyg-11 family member B, cell cycle regulator	No	NaN	0	0	0	0
19254	ZYX	zyxin	No	NaN	0	0	0	0
19255	ZZEF1	zinc finger ZZ-type and EF-hand domain contain...	Yes	Multiple Disorders	3	1	1	1
19256	ZZZ3	zinc finger ZZ-type containing 3	No	NaN	0	0	0	0

Fig. 3: Gene list with risk category - Bottom 10 Sample View

Fig. 4: Gene Association Metrics

2.3 Methodology

Data Wrangling: Used data Wrangling techniques to integrate the multi-omics data into a single data matrix.

Model Building: Utilized Explainable AI frameworks like t-SNE (t-Distributed Neighbour Embedding) to build and interpret the predictive model. And, other models like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) has been refereed.

Feature Importance: Used XAI techniques to rank features based on their importance for prediction. This could be RAF (Risk allele frequency), Number of associations, gene mutations, expression levels, or specific proteins that are key in classifications.

Clinical Interpretation: Translated these feature importance into useful insights. For example, if a certain gene's mutation is highly ranked, it could be a potential drug target in future.

2.4 Model Explainability

In this model, to differentiate the potential vulnerable risk gene, TSNE method (t- distributed stochastic neighbour embedding method) has been employed with necessary perplexity value. Perplexity value is a tuneable parameter which considers the nearest neighbours and helps to balances between local and global aspects. It helps us to visually explore, to identify and to interpret the risk clusters from the high dimensional datasets. For each gene in high dimensional space, t-SNE looks at its neighbours and measures how close they are.

Figure colored by risk status (Yellow - Risk Associations and Blue - No Risk Associations) and shape by three categories.

To Interpret this data requires a deep understanding of molecular biology, and the specific diseases in question. And, this kind of research can be crucial for the development of new therapies and understanding the underlying mechanisms of complex mental health conditions.

For example: 1. Genetic association studies, which might show correlations between variants of the certain gene and the incidence of neurological disorders. (gene like NRG1, DISK1

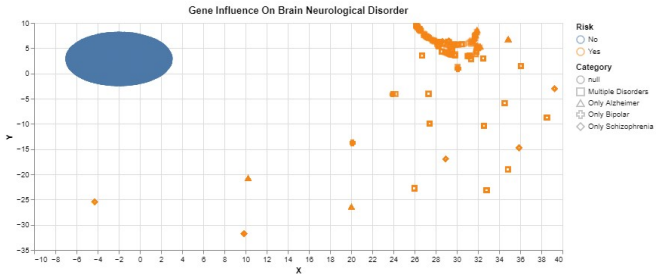

Fig. 5: TSNE with perplexity=5

which has been identified as a risk category) 2. Expression studies, which would examine whether the expression of certain gene (like NRG1) is altered in the brains of individuals with these conditions. 3. Functional studies, to understand what NRG1 does in the brain and how mutations might impact.

Few other factors/features which can also be considered in future studies: 1. Gene Length: The number of base pairs in the gene. The longer genes might have more regulatory regions and potentially more complexity. 2. Number of Exons, 3. GC Content, 4. Mutation Rate, 5. Tissue Specificity, 6. Evolutionary Conservation Score, 7. Functional Annotation Score, 8. Transcription Factor Binding Sites, 9. RNA Splice Sites, 10. 3' UTR and 5' UTR Lengths, 11. Synonymous/Non-synonymous Mutation Ratio, 12. Network Centrality Measures.

3 Conclusion

Finally conclude that gene data sets were collected from GWAS Catalog (Genome wide associations study) and three main neurological conditions has been identified and focused such us Alzheimer, Bi-Polar and Schizophrenia. Also, In this paper various risk and associated genes were studied and summarised using XAI techniques. Even though, identifying the exact root cause is extremely difficult for these severe disorders, currently using explainable AI able to highlight the key genes, pathways, or molecular mechanisms that differentiate the potential vulnerable.

This report supporting materials, coding has been documented here in this Github Page. (https://github.com/Vivekanandr/Multi-Omics-Analysis)

In future, new studies from various databases can also be employed with this paper where a set of biologically interpretable, actionable insights that healthcare providers and researchers can follow up on.

References:

1.The genetic relationships between brain structure and schizophrenia [**https://doi.org/10.1038/s41467-023-43567-7**]

2.BioMOBS: A multi-omics visual analytics workflow for biomolecular insight generation. [**https://doi.org/10.1371/journal.pone.0295361**]

3.Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder: [**https://www.science.org/doi/10.1126/science.aat8127**]

4.The NHGRI-EBI Catalog of human genome-wide association studies (https://www.ebi.ac.uk/gwas/)

5.NIMH Data and Genome Archive: https://www.nimhgenetics.org/download-tool/SZ

[1]

[1]This is a sample report, educational, non-commercial and for demonstration purpose only