Machine Learning-Based Identification of Narcolepsy Predictive Genes

Submitted To

Dr. Meenal Kowshik

BITS Pilani

Name	BITS ID
Gaurang Aswal	2020B1A71960G
Jimit Shah	2020B1A72097G
Laukik B Nakhwa	2020B1A81932G

Birla Institute of Technology and Science, KK Birla Goa Campus, India

Introduction

Narcolepsy, a complex neurological disorder characterized by excessive daytime sleepiness, sudden loss of muscle tone (cataplexy), hallucinations, and disrupted nocturnal sleep patterns, poses significant challenges to both affected individuals and healthcare providers. While relatively rare, with an estimated prevalence of 25 to 50 per 100,000 individuals worldwide, narcolepsy exerts a profound impact on daily functioning, social interactions, and overall quality of life (American Academy of Sleep Medicine, 2014).

The manifestations of narcolepsy extend beyond mere somnolence, as individuals may experience cataplexy, a sudden and temporary loss of muscle tone typically triggered by strong emotions such as laughter, surprise, or anger (American Academy of Sleep Medicine, 2014). Furthermore, individuals with narcolepsy may encounter vivid hallucinations upon falling asleep or waking up, alongside episodes of sleep paralysis, wherein they remain temporarily unable to move or speak upon awakening (American Academy of Sleep Medicine, 2014).

Narcolepsy is characterised by dysfunction in regulating sleep-wake cycles, with disruptions in the rapid eye movement (REM) phase of sleep being particularly prominent. This dysregulation often leads to fragmented nocturnal sleep and the intrusion of REM sleep phenomena into wakefulness, contributing to the constellation of symptoms observed in affected individuals (Ponz et al., 2015).

Genetic factors play a pivotal role in the aetiology of narcolepsy, with a strong association observed with certain human leukocyte antigen (HLA) class II genes, specifically HLA-DQB1*06:02. However, the precise mechanisms underlying narcolepsy susceptibility remain incompletely understood, suggesting the involvement of additional genetic loci and environmental factors in disease pathogenesis (Mignot et al., 2001).

In recent years, advancements in machine learning (ML) techniques have offered promising avenues for unravelling the intricate genetic underpinnings of complex diseases such as narcolepsy. ML algorithms can analyse vast datasets comprising genetic, clinical, and demographic information to identify patterns, correlations, and predictive models that elude traditional analytical approaches (Zou et al., 2019).

In this report, we leverage the power of ML methodologies to predict narcolepsy and uncover crucial genes implicated in disease susceptibility. By integrating genomic data from diverse sources and employing sophisticated ML algorithms, we aim to elucidate the genetic architecture of narcolepsy, paving the way for improved diagnostic strategies and targeted therapeutic interventions. Through a comprehensive analysis of narcolepsy-associated genes and their functional implications, we endeavour to enhance our understanding of this debilitating disorder and facilitate the development of personalised approaches for disease management.

Methodology

Data Retrieval and Preprocessing:

We began by extracting gene expression data from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) database for the study accession number GSE21592 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21592) (Barrett et al., 2013). This dataset comprises transcriptomic profiles associated with the target disease condition. This case-control study was performed comparing 10 narcoleptic patients (both sexes, mean age 50, median age 50) with 10 age- and sex-matched healthy controls. We filtered and normalised the raw expression data through rigorous data preprocessing steps to ensure consistency and reliability in subsequent analyses. GSE21592_RAW.tar supplementary resource was downloaded.
Training
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69371/

Validation
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69370/

Identification of Differentially Expressed Genes (DEGs) and Autoimmune-Related Genes (ARGs):

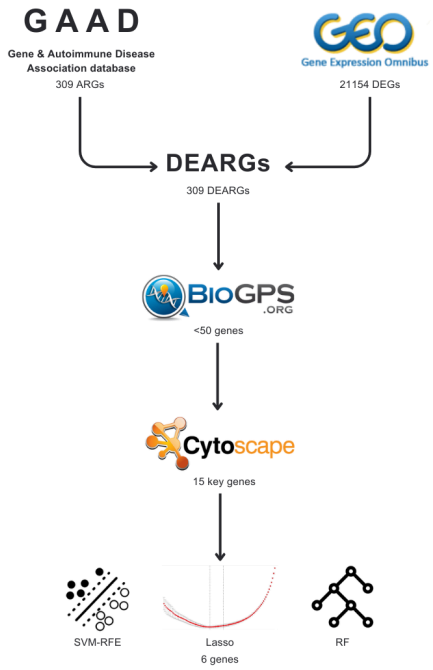
(DEGs refer to genes whose expression levels significantly differ between different conditions or samples, regardless of whether they are related to autoimmune diseases. In contrast, ARGs specifically refer to genes implicated in autoimmune diseases, focusing on their role in autoimmunity)

Next, we conducted a comprehensive analysis to identify differentially expressed genes (DEGs) using established statistical methods. Specifically, we employed differential expression analysis techniques to compare gene expression levels between disease samples and control samples, extracting a total of 21,154 DEGs from the GSE21592 dataset.

Simultaneously, we curated a list of autoimmune-related genes (ARGs) sourced from the Gene and Autoimmune Disease Association Database (GAAD) (https://gaad.medgenius.info/int)(Zhang & Wiemann, 2009). This repository contains a comprehensive compilation of genes implicated in autoimmune disorders. From the GAAD database, we extracted a total of 309 ARGs relevant to our study.

Integration and Selection of Differentially Expressed Autoimmune-Related Genes (DEARGs):

To delineate the intersection between DEGs and ARGs, we conducted a systematic comparison to identify differentially expressed autoimmune-related genes (DEARGs). By overlaying the lists of DEGs and ARGs, we identified 309 DEARGs that exhibit dysregulated expression patterns specifically associated with the target autoimmune condition.



Identified Top 10 DEARGs refer to extracted data.xlsx as per ascending P-Values.

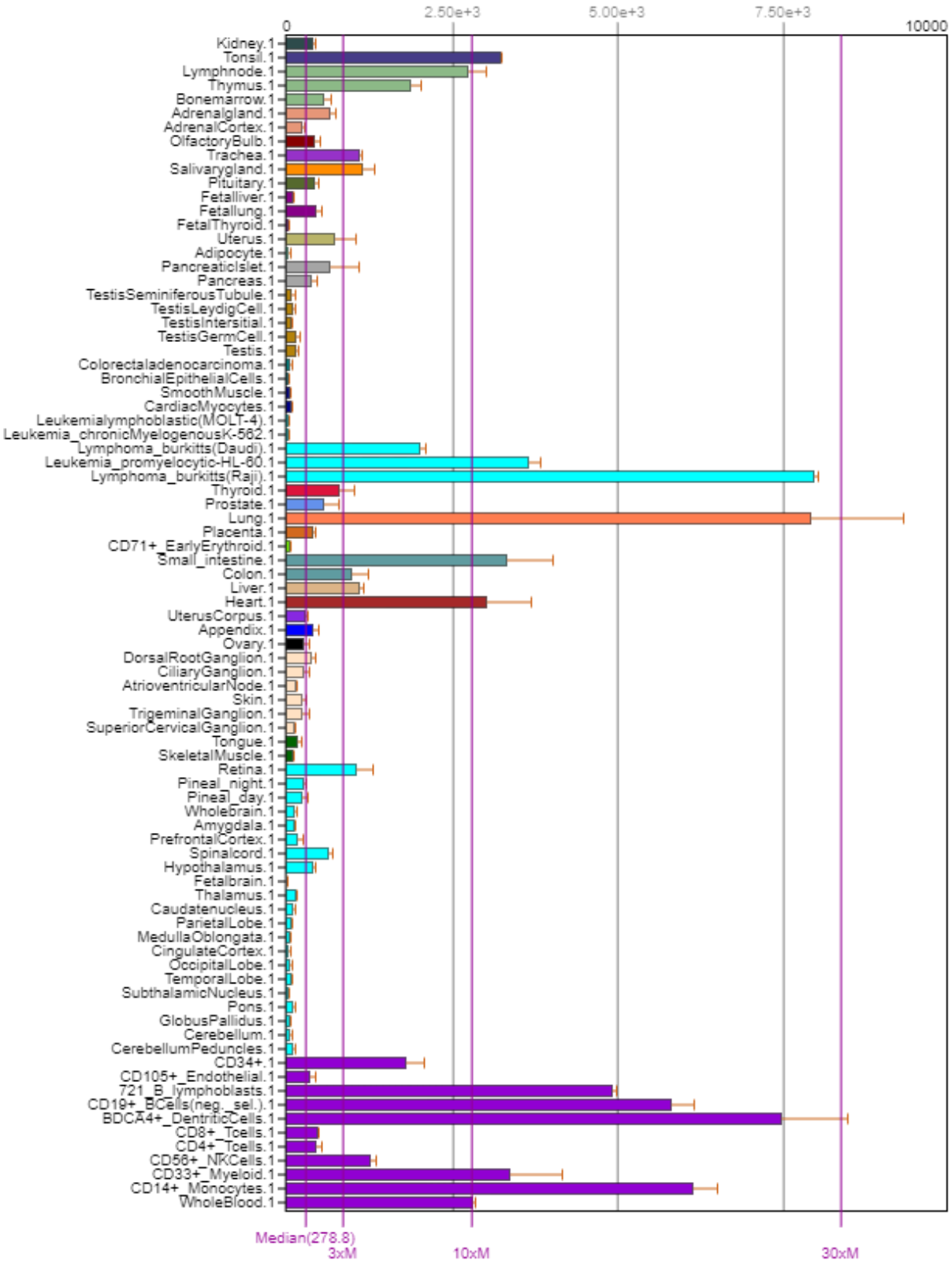
Gene
HLA-DQB1
HLA-DQB1
HLA-DQB1
MOG
MOG
CTSH
CHKB
TAC1
PENK
HLA-DQA1

Evaluation of Gene Expression Levels and Relevance to Disease Pathology:

To further refine our candidate gene pool, we leveraged the BioGPS platform—an online resource for exploring gene expression profiles across diverse tissues and cell types. Through

BioGPS, we assessed the expression levels of all DEARGs and prioritised genes demonstrating relevance to the disease phenotype. We curated a subset of approximately <50 genes based on their tissue-specific expression patterns and known associations with disease pathology.

Data for HLA-DQB1 tissue-specific expression :


0
2.50e+3
5.00e+3
7.50e+3
10000
Kidney.1
Tonsil.1
Lymphnode.1
Thymus.1
Bonemarrow.1
Adrenalgland.1
AdrenalCortex.1
OlfactoryBulb.1
Trachea.1
Salivarygland.1
Pituitary.1
Fetalliver.1
Fetallung.1
FetalThyroid.1
Uterus.1
Adipocyte.1
PancreaticIslet.1
Pancreas.1
TestisSeminiferousTubule.1
TestisLeydigCell.1
TestisInterstitial.1
TestisGermCell.1
Testis.1
Colorectaladenocarcinoma.1
BronchialEpithelialCells.1
SmoothMuscle.1
CardiacMyocytes.1
Leukemialymphoblastic(MOLT-4).1
Leukemia_chronicMyelogenousK-562.1
Lymphoma_burkitts(Daudi).1
Leukemia_promyelocytic-HL-60.1
Lymphoma_burkitts(Raji).1
Thyroid.1
Prostate.1
Lung.1
Placenta.1
CD71+_EarlyErythroid.1
Small_intestine.1
Colon.1
Liver.1
Heart.1
UterusCorpus.1
Appendix.1
Ovary.1
DorsalRootGanglion.1
CiliaryGanglion.1
AtrioventricularNode.1
Skin.1
TrigeminalGanglion.1
SuperiorCervicalGanglion.1
Tongue.1
SkeletalMuscle.1
Retina.1
Pineal_night.1
Pineal_day.1
Wholebrain.1
Amygdala.1
PrefrontalCortex.1
Spinalcord.1
Hypothalamus.1
Fetalbrain.1
Thalamus.1
Caudatenucleus.1
ParietalLobe.1
MedullaOblongata.1
CingulateCortex.1
OccipitalLobe.1
TemporalLobe.1
SubthalamicNucleus.1
Pons.1
GlobusPallidus.1
Cerebellum.1
CerebellumPeduncles.1
CD34+.1
CD105+_Endothelial.1
721_B_lymphoblasts.1
CD19+_BCells(neg._sel.).1
BDCA4+_DentriticCells.1
CD8+_Tcells.1
CD4+_Tcells.1
CD56+_NKCells.1
CD33+_Myeloid.1
CD14+_Monocytes.1
WholeBlood.1
Median(278.8)
3xM
10xM
30xM

Prioritisation of Candidate Genes:

Subsequently, we employed stringent statistical criteria to prioritise candidate genes for further analysis. By assessing the significance of differential expression using p-values derived from appropriate statistical tests, we selected the top 10-20 genes exhibiting the most robust associations with the disease phenotype.

Integration of Machine Learning Models for Gene Selection:

To enhance the precision and accuracy of our gene selection process, we tried to employ advanced machine learning algorithms, including Support Vector Machine-Recursive Feature Elimination (SVM-RFE), Lasso logistic regression, and Random Forest (RF) (Tibshirani, 1996). These algorithms should facilitate the identification of a refined subset of genes with the highest predictive power for the target autoimmune disease.

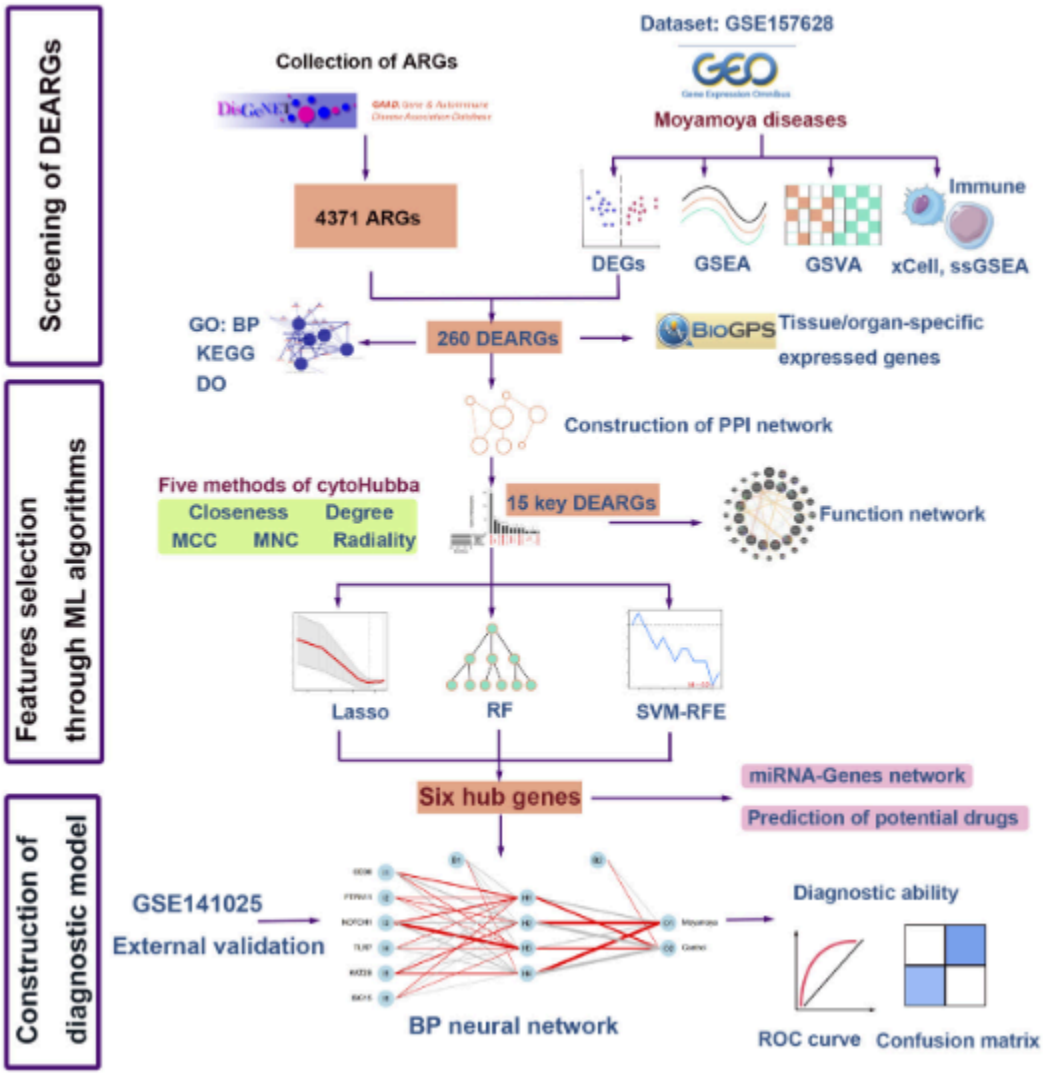
Supporting Code

Please refer to the Github Repo Link with all code and supporting data files.

Further Approach

Predictive Modeling and Finalization of Important Genes:

In the future, we can utilise the selected gene panel to construct predictive models aimed at discerning disease status based on gene expression profiles. By integrating machine learning-derived gene signatures, we could generate a comprehensive predictive framework capable of accurately identifying individuals at risk of the autoimmune condition. Thus, through the amalgamation of bioinformatics analyses and machine learning methodologies, we can identify and validate a set of pivotal genes crucial for the pathogenesis and diagnosis of the target autoimmune disease.



References

1. GAAD website : https://gaad.medgenius.info/int
2. Lu, G., Hao, X., Chen, W.-H., & Mu, S. (2018). Gaad: A gene and Autoimmiune Disease Association Database. *Genomics, Proteomics & Bioinformatics*, *16*(4), 252–261. doi:10.1016/j.gpb.2018.05.001
3. Li, S., Han, Y., Zhang, Q., Tang, D., Li, J., & Weng, L. (2022). Comprehensive molecular analyses of an autoimmune-related gene predictive model and immune infiltrations using machine learning methods in moyamoya disease. *Frontiers in Molecular Biosciences*, *9*. doi:10.3389/fmolb.2022.991425
4. American Academy of Sleep Medicine. (2014). International Classification of Sleep Disorders - Third Edition (ICSD-3). Darien, IL: American Academy of Sleep Medicine.
5. Mignot, E., Lin, L., Rogers, W., Honda, Y., Qiu, X., Lin, X., ... Okun, M. (2001). Complex HLA-DR and -DQ Interactions Confer Risk of Narcolepsy-Cataplexy in Three Ethnic Groups. American Journal of Human Genetics, 68(3), 686–699. https://doi.org/10.1086/318798
6. Ponz, A., Khatami, R., Poryazova, R., Werth, E., Boesiger, P., Schwartz, S., ... Bassetti, C. L. (2015). Abnormal activity in reward brain circuits in human narcolepsy with cataplexy. Annals of Neurology, 77(4), 537–547. https://doi.org/10.1002/ana.24350
7. Zou, Q., Qu, K., Luo, Y., Yin, D., & Ju, Y. (2019). Toward machine learning application in sleep medicine: A comprehensive review. Sleep Medicine Reviews, 46, 27–40. https://doi.org/10.1016/j.smrv.2019.03.004
8. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research, 41(D1), D991-D995.
9. Zhang, J. D., & Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics, 25(11), 1470-1471.
10. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.