"It's All in Your Head!": Identifying Potential Biomarkers for Bipolar Disorder, Schizophrenia, and Major Depressive Disorder from the Gene Expression Data of Postmortem Human Dorsolateral Prefrontal Cortex

Jose Marie C. Cordova[1], Alex C. Gonzaga, PhD[1], and Joey Mark S. Diaz, PhD[1]

[1] University of the Philippines Manila, Manila 1000, Philippines

Abstract. The underlying biology of mental health disorders, such as major depressive disorder (MDD), bipolar disorder (BD), and schizophrenia (SCZ), remains an enigma for healthcare professionals and biomedical researchers. This study investigates the genes associated with these disorders, focusing on the dorsolateral prefrontal cortex (DLPFC), which is established as a key regulator of emotions. Utilizing the DLPFC gene expression dataset from the Gene Expression Omnibus (GSE208338), the data underwent normalization, dimension reduction, and various classification machine learning techniques. Models with an area under the curve (AUC) of 70% or higher were considered valid. Fifty-eight features were extracted from multinomial models, and forty-four from binomial models, with annotations obtained through Ensembl Biomart. Of these, six features were annotated, linking genes such as NARS1, SEPTIN2, PSMB2, ANAPC5, and RNU1-124P to neurological disorders. NARS1 and PSMB2, in particular, were associated with neurological conditions. An in-depth analysis on the unannotated features would help uncover the disorder. These findings could enhance the early detection, accurate diagnosis, and treatment of BD, MDD, and SCZ, potentially reducing the financial burden of trial-and-error treatments and improving patients' quality of life. Furthermore, this research contributes to reducing the stigma by highlighting the biological underpinnings of these mental health disorders.

Keywords: Mental Health, Dorsolateral Prefrontal Cortex. Machine Learning, Depression, Biomarkers, Explainable AI

1 Introduction

Mental health disorders constitute approximately 16.5% of the global disease burden. [1] According to a 2019 survey, 970 million individuals worldwide are affected by mental health disorders, with major depressive disorder (MDD) being the most prevalent, affecting 280 million people. [2] Additionally, 40 million individuals have bipolar disorder (BD), and 24 million suffer from schizophrenia (SCZ). [2] In a meta-study examining the mortality of mental health patients, a total of 338,381 deaths

J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024) , Atlantis Highlights in Computer Sciences 23 ,
https://doi.org/10.2991/978-94-6463-684-0_13

were recorded, with 67.3% attributed to natural causes, 17.5% to unnatural causes, and the remainder due to unspecified causes. [3] In the Asia-Pacific region, the most common mental health disorders include depressive disorders, anxiety disorders, post-traumatic stress disorder (PTSD), suicide, and substance abuse disorders. [4] The presence of a mental health disorder significantly increases an individual's mortality risk and their years of potential life lost (YPLL) compared to the general population. [5] The YPLL metric is employed to quantify social and economic losses due to premature death, emphasizing specific causes of death affecting younger age groups. [5]

While interventions for mental health disorders have proven effective, the need for cost-effective solutions remains pressing. [4] Furthermore, the development and implementation of early detection systems are critical. [4] Genetic biomarkers have indicated that certain populations are more susceptible to these disorders. [3,6,7] However, there is limited research on the correlation between genes in the dorsolateral prefrontal cortex (DLPFC), the brain's emotional control center, and disorders such as MDD, BD, and SCZ. [8] The DLPFC is involved in functions such as planning, risk-taking, motivation, and emotional biases. [9] This study aimed to identify gene expression biomarkers or features in the DLPFC that are specific to each disorder as well as those common across MDD, BD, and SCZ. It employed various machine learning (ML) techniques to achieve optimal predictive outcomes.

2 Materials and Methods

2.1 Data Source

The dataset with identification GSE208338 came from querying the phrase “human[organism] AND depression” in Gene Expression Omnibus. It is titled “Expression data from postmortem human dorsolateral prefrontal cortex - psychiatric disorders & healthy controls” (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208338) which was publicly available last November 17, 2022.

The dataset contained 169 CEL files which represent 169 adults aged 18 to 87 years old. These files contain probe intensity data of DLPFC transcripts which were hybridized using Affymetrix Human Exon 1.0 ST v2 arrays. Of the 169 samples, 68 were diagnosed with SCZ, 15 with BD, 24 suffered from MDD, and the remaining 62 were control subjects without any of the mentioned disorders.

2.2 Methodology

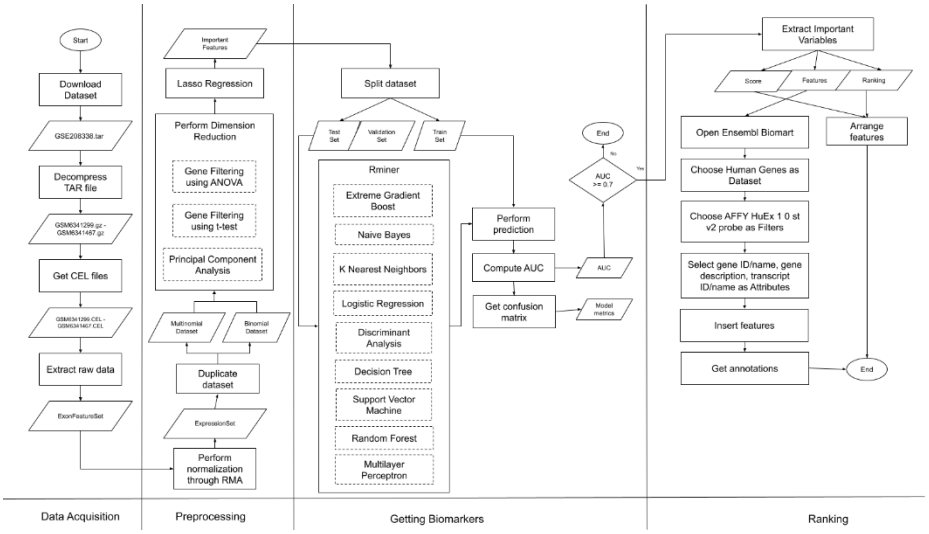

Fig. 1. Process flow diagram of the study

Data Acquisition. The dataset from 2.1 was downloaded and extracted. Using the package oligo, the Affymetrix CEL files were read in R. The gene probe information, in the datatype of ExonFeatureSet, was stored.

Preprocessing. To convert the probe information into gene expression data, background correction, normalization, through the use of oligo's Robust Multichip Average (RMA), were performed. [10] The ExpressionSet datatype that was returned by this function included a matrix with 22,011 features from 167 individuals, each named by their CEL filename and their Affymetrix GeneChipTM Human Exon 1.0 ST Array v2 na36 hg19 probe IDs. A feature scal-ing step was performed to prevent the dataset's values from being skewed. This included standardizing the dataset to have a normal distribution with a mean of 0 and a standard deviation of 1.

Due to the matrix being duplicated, one catered to the binomial and the other to the multinomial. BPD = 1, MDD = 2, SCZ = 3, and control (CTL) = 9 made up the multinomial diagnostic column. However, it was just CTL = 9 and with disorder (NCTL) = 1 for binomial.

Dimension Reduction. In this dataset, dimension reduction is beneficial because it eliminates noise and redundancy, which reduces computation time while enhancing model performance and accuracy. [11] To minimize the processing time and effort, a parallel dimen- sion reduction using gene filtering (GF) and principal component analysis (PCA) was done to ExpressionSet.

GF included normality and t-test (normal and significantly different), analysis of variance (ANOVA), and coefficient of variation (cov), taking only cov < 0.2, which indicates both the experimental effect and the measurement precision are large. The elbow approach was used to extract nine principal components (PC) from the PCA. Interpreting each primary component is challenging because this study is discovery-driven. To obtain the best-performing features across the nine PCs, feature extraction was carried out in this instance utilizing PCATools' plotloadings() function in conjunction with the getComponents() function. The chosen characteristics from PCA and GF were integrated. Model regularization was used in the combined dataset to prevent overfitting.

Multicollinearity was first examined. When calculating the variance inflation factor (VIF <= 5), only features with little to no collinearity will be kept. The features "2908474" and "2908505" from the multinomial dataset were eliminated from the list in this stage because they have perfect collinearity with one another. Furthermore, Rminer employed the feature selection process known as sensitivity analysis backward selection (SABS). The model in SABS fits every feature, then eliminates each feature one at a time until only statistically significant features remain. [12] This procedure removed the covariates with the greatest p-value for the subsequent iteration after taking them into account for each one.

There were 37 features from GF, 23 from PCA, and 58 features from combined for multinomial models (CTL vs. MDD vs. SCZ vs. BD).

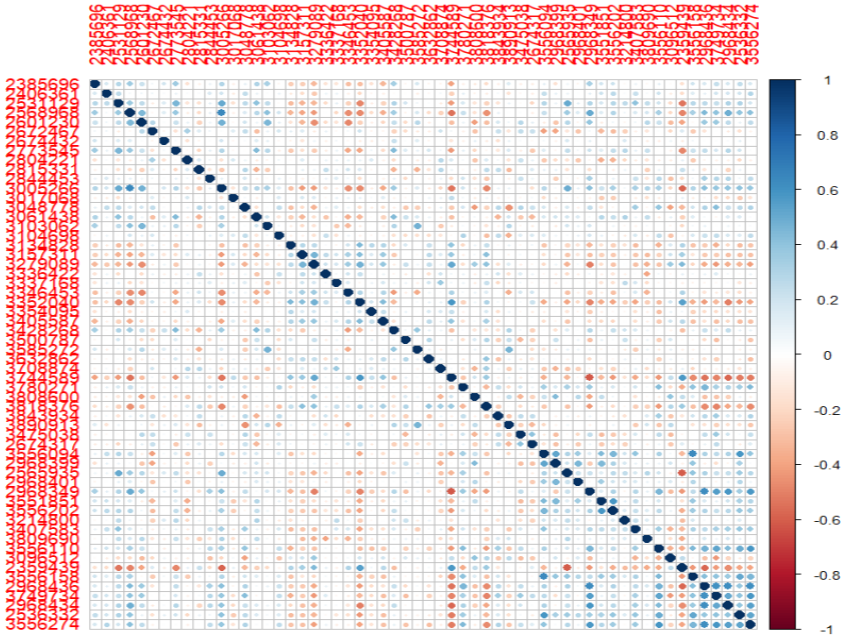

Fig. 2. Correlation matrix of each selected feature for multinomial models

There were 33 features from GF, 52 from PCA, and 44 from combined for binomial models (CTL vs. non-control NCTL).

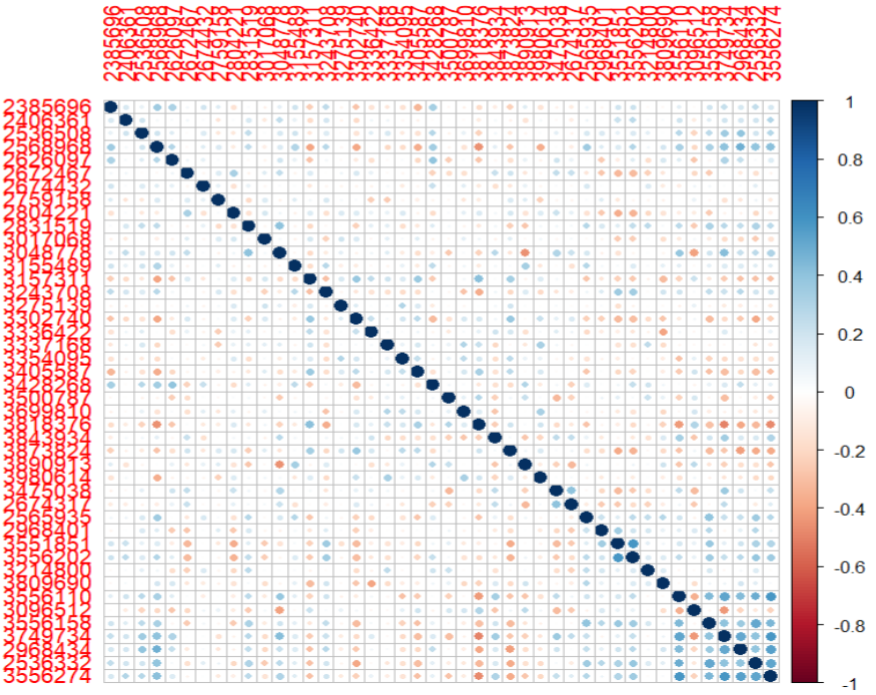

Fig. 3. The correlation matrix of each selected feature for binomial models

The cross-validation method used in this investigation was the train-validation-test. Of the dataset, 75% was utilized for training and 25% was used for testing. To guarantee that the diagnosis was dispersed equally throughout the set, the training dataset for the multinomial model was upsampled to include 25% BD, 25% MDD, 25% SCZ, and 25% CTL. For the binomial model, 50% CTL and 50% NCTL are utilized. The training set was divided into a 75% training set and a 25% validation set. The validation set was used to adjust the parameters, while the test set was used to forecast the model.

Data Processing. The following machine learning models were employed: Random Forest (RF), Extreme Gradient Boost (XGB), Decision Tree (DT), Conditional Inference Tree (CIT), K Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Discriminant Analysis (DA), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The "fit" function in the Rminer package automatically tested each of the previously described models and selected the top-performing ML by

obtaining the highest AUC. Rminer also automatically adjusted the hyperparameters using the "mparheuristic" function.

Rminer was used to fit the training set against each designated model. It calculated the AUC for each model. Additionally, it assigns a descending order to the models based on their AUC.

The model with the highest AUC was selected during model fitting. Since there isn't a "gold standard" for BD, MDD, and SCZ biomarkers, Area Under Curve AUC can be used as an evaluation statistic to gauge how accurate the model is. [11] A model that has an AUC of 70% or higher was retained. The minimum level for what constitutes "acceptable discrimination" is 70%. [13] The model's performance measures, including accuracy, sensitivity, specificity, precision, and f-score, were then calculated once the confusion matrix was created.

Annotation. The annotations were obtained using Ensembl BioMart. AFFY HuEx 1.0 st v2 probe IDs for filters, human genes (GRCh38.p14) for the dataset, and finally, gene stable ID, gene stable ID version, transcript stable ID, transcript stable ID version, AFFY HuEx 1.0 st v2 probe, gene description, gene name, and transcript name for the attributes were the parameters.

Ranking. Using the metric of average absolute departure from the median, each top-performing model in binomial and multinomial datasets was subjected to Rminer's Importance using one-dimensional sensitivity analysis. This procedure identified features that demonstrated their significance in the prediction. The models, SVM for binomial datasets and RF for multinomial datasets, were used to extract the significant features and their relative relevance.

3 Results

Finding biomarkers that can differentiate BD, MDD, and SCZ from a collection of gene expression datasets is one of the goals of this research. Each of the models mentioned in the data processing stage was fitted by the features that were extracted from the preprocessing step of the binomial and multinomial datasets. The AUC and feature importance of the two models were taken into account.

3.1 Binomial Dataset

SVM is the best model to use for this dataset with an AUC of 71%. It has an "acceptable discrimination". [11]

The model's accuracy was 70.7317%, precision was 81.8181%, recall was 69.2308%, and F1 score was 75%. The radial basis/gaussian "rbfdot" kernel, which cost 0.0170484, was the SVM hyperparameter that produced the best prediction.

Table 1. Ranking of the features extracted from SVM for BN

Rank	Features	BN	Rank	Features	BN
1	2385696	2831519	24	3699810	
2	2406361	3302740	25	3818376	
3	2536508	3354095	26	3843934	
4	2568968	3155489	27	3873824	
5	2626097	2804221	28	3890913	
6	2672467	3843934	29	3980614	
7	2674432	2406361	30	3475038	
8	2759158	2672467	31	2674317	
9	2804221	3500787	32	2565935	
10	2831519	3749734	33	2968401	
11	3017068	2536508	34	3551851	
12	3048778	3017068	35	3556202	
13	3155489	3980614	36	3214800	
14	3157311	2674432	37	3809690	
15	3243708	3337168	38	3556110	
16	3275139	3096512	39	3096512	
17	3302740	3048778	40	3556158	
18	3336422	3699810	41	3749734	
19	3337168	2565935	42	2968434	
20	3354095	3243708	43	2536332	
21	3405587		44	3556274	
22	3428268				
23	3500787				

For the binomial dataset, of the original 44 features, 23 were retained for BN. Table 1 lists all the extracted features together with the features kept by BN.

3.2 Multinomial Dataset

RF is the best model for this dataset with an AUC of 73%. The AUC is considered as "acceptable discrimination". [11]

500 trees with six variable trials per split was the RF hyperparameter that produced the best prediction. With an F1 score of 0%, precision of 0%, and recall of 0%, the model can predict BD 90.2439% with accuracy. With an accuracy of 80.4878%, precision of 33.3333%, recall of 33.3333%, and F1 score of 33.3333%, this model is capable of predicting MDD. For SCZ, the model's accuracy was 70.7317%. has an F1 score of 62.5%, recall of 58.8235%, and precision of 66.6667%. The model's accuracy, precision, recall, and F1 score for CTL are 78.0488%, 68.75%, and 73.3333%, respectively. With a weighted accuracy of 75.1338%, the overall accuracy is 58.5366%.

Table 2. Ranking of the features extracted from MN

Rank	Features	MN	Rank	Features	MN
1	2385696	3157311	30	3632862	2674317
2	2406361	3744589	31	3708874	3556110
3	2531129	3352040	32	3744589	3556202
4	2568968	2844453	33	3780271	3708874
5	2601230	3632862	34	3808600	2531129
6	2672467	3134828	35	3818376	3818376
7	2674432	3500787	36	3843934	3061438
8	2773545	2773545	37	3890913	3103062
9	2804221	2406361	38	3475038	3104698
10	2815331	3890913	39	2674317	3048778
11	2844453	2968401	40	3556094	3428268
12	3005266	3005266	41	2968399	3556274
13	3017068	3354095	42	2565935	3407583
14	3048778	3475038	43	2968401	2968349
15	3061438	3337168	44	2968349	3017068
16	3103062	3346453	45	3551851	2536332
17	3104698	2568968	46	3556202	3551851
18	3134828	2674432	47	3214800	
19	3157311	2601230	48	3407583	
20	3279089	3405587	49	3809690	
21	3336422	3585272	50	3556110	
22	3337168	3780271	51	3096512	
23	3346453	3556094	52	2359439	
24	3352040	3809690	53	3556158	
25	3354095	2385696	54	2968436	
26	3405587	2359439	55	3749734	
27	3428268	3808600	56	2968434	
28	3500787	3556158	57	2536332	
29	3585272	3336422	58	3556274	

For MN, of the original 58 features, only 46 remained. Table 2 lists all the extracted features together with the features kept by MN.

3.3 Top Performing Features

Table 3. Top 20 features from the BSVE and MN

Rank	BN	MN	Rank	BN	MN	Rank	BN	MN

1	3354095	2831519	8	2672467	2773545	15	3337168	3337168
2	3302740	3302740	9	3500787	2406361	16	3096512	3346453
3	2831519	3354095	10	3749734	3890913	17	3048778	2568968
4	3428268	3155489	11	2536508	2968401	18	3699810	2674432
5	2804221	2804221	12	3017068	3005266	19	2565935	2601230
6	3500787	3843934	13	3980614	3354095	20	3243708	3405587
7	3843934	2406361	14	2674432	3475038			

BN's top 20 features were "2831519", "3302740", "3354095", "3155489", "2804221", "3843934", "2406361", "2672467", "3500787", "3749734", "2536508", "3017068", "3980614", "2674432", "3337168", "3096512", "3048778", "3699810", "2565935", and "3243708". Table 3 shows the 20 most important features of both BN and MN.

For MN, these 20 features are important for predicting BD: "2385696", "2406361", "2531129", "2568968", "2601230", "2674432", "2773545", "2844453", "3005266", "3017068", "3048778", "3061438", "3103062", "3104698", "3134828", "3157311", "3336422", "3337168", "3346453", and "3352040". For MDD, these 20 features are "2385696", "2406361", "2531129", "2568968", "2601230", "2674432", "2773545", "2844453", "3005266", "3017068", "3048778", "3061438", "3103062", "3104698", "3134828", "3157311", "3336422", "3337168", "3346453", and "3352040". For SCZ, these 20 features are "2385696", "2406361", "2531129", "2568968", "2601230", "2674432", "2773545", "2844453", "3005266", "3017068", "3048778", "3061438", "3103062", "3104698", "3134828", "3157311", "3336422", "3337168", "3346453", and "3352040".

Additionally, the features common on both BN and MN are "2831519", "3302740", "3354095", "2804221", "3843934", "2406361", "2406361", "3354095", "3337168", and "2674432".

3.4 Biological Processes related to BD, MDD, and SCZ

Ensembl Biomart was compared to the features mentioned in Tables 1 and 2. Only four of the 58 probes for multinomials ("2406361", "3475038", "3096512" "2536332) had annotations, whereas five of the 44 probes for binomial models ("2406361", "2536508", "3475038", "3096512" "2536332) had annotations. The annotations were in Table 4.4.1. Each gene was compared to Gene Cards, a website that enables researchers to collect data on the genes' locations, phenotypes, diseases, and other characteristics, in order to obtain relevant information about the genes.

Table 4. Ensembl Biomart annotations per probe

Probe ID	Gene stable ID	Gene name	Transcript stable ID	Transcript name	Exon stable ID

3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00001225301
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003650295
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003482714
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003580355
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003461119
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003461949
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003642036
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003567090
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003556619
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003478089
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003509365
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00003512665
3809690	ENSG00000134440	NARS1	ENST00000256854	NARS1-201	ENSE00001625791
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003650295
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003482714
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003664208
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003569994
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003653246
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003691817
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003480632
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003671257
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003491151
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003524260
3809690	ENSG00000134440	NARS1	ENST00000586807	NARS1-204	ENSE00003459130
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00002216753
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00003682603
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00003551779
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00003679217
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00003593533
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00003515559
3809690	ENSG00000134440	NARS1	ENST00000411676	NARS1-202	ENSE00001782037
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003650295
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003482714
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003580355
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00002216503
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003487383
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00003688039
3809690	ENSG00000134440	NARS1	ENST00000540592	NARS1-203	ENSE00001655265
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00003580355
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00003461119
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00003461949

3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00002840453
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00003617558
3809690	ENSG00000134440	NARS1	ENST00000588661	NARS1-208	ENSE00002767343
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00003580355
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00003461119
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00003461949
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00003617558
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00002767343
3809690	ENSG00000134440	NARS1	ENST00000587675	NARS1-207	ENSE00002895921
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00003650295
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00003482714
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00003631659
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00003580355
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00002216503
3809690	ENSG00000134440	NARS1	ENST00000587194	NARS1-205	ENSE00002917156
3809690	ENSG00000134440	NARS1	ENST00000590123	NARS1-211	ENSE00003551779
3809690	ENSG00000134440	NARS1	ENST00000590123	NARS1-211	ENSE00002886021
3809690	ENSG00000134440	NARS1	ENST00000590123	NARS1-211	ENSE00002952061
3809690	ENSG00000134440	NARS1	ENST00000590123	NARS1-211	ENSE00002910158
3809690	ENSG00000134440	NARS1	ENST00000590123	NARS1-211	ENSE00002783833
3809690	ENSG00000134440	NARS1	ENST00000587366	NARS1-206	ENSE00002823664
3809690	ENSG00000134440	NARS1	ENST00000587366	NARS1-206	ENSE00003627939
3809690	ENSG00000134440	NARS1	ENST00000587366	NARS1-206	ENSE00002751004
3809690	ENSG00000134440	NARS1	ENST00000587366	NARS1-206	ENSE00002768591
2406361	ENSG00000126067	PSMB2	ENST00000630477	PSMB2-203	ENSE00003732167
2406361	ENSG00000126067	PSMB2	ENST00000630477	PSMB2-203	ENSE00003740092
2406361	ENSG00000126067	PSMB2	ENST00000630477	PSMB2-203	ENSE00003773632
2406361	ENSG00000126067	PSMB2	ENST00000630477	PSMB2-203	ENSE00003771043
2406361	ENSG00000126067	PSMB2	ENST00000630477	PSMB2-203	ENSE00003766013
2406361	ENSG00000126067	PSMB2	ENST00000621781	PSMB2-202	ENSE00001124812
2406361	ENSG00000126067	PSMB2	ENST00000621781	PSMB2-202	ENSE00003732167
2406361	ENSG00000126067	PSMB2	ENST00000621781	PSMB2-202	ENSE00003740092
2406361	ENSG00000126067	PSMB2	ENST00000621781	PSMB2-202	ENSE00003751487
2406361	ENSG00000126067	PSMB2	ENST00000621781	PSMB2-202	ENSE00003740603
3096512	ENSG00000200731	RNU1-124P	ENST00000363861	RNU1-124P-201	ENSE00001438624
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003462509
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003464499
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003674994
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003642882
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003494282
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003574156
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003546899
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00002218588
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00003686084

3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00002229210
3475038	ENSG00000089053	ANAPC5	ENST00000545218	ANAPC5-221	ENSE00002323683
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00003679843
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00003646772
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00001459997
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00001555489
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00001559929
2536332	ENSG00000168385	SEPTIN2	ENST00000428524	SEPTIN2-214	ENSE00001652822
2536332	ENSG00000168385	SEPTIN2	ENST00000445030	SEPTIN2-221	ENSE00003679843
2536332	ENSG00000168385	SEPTIN2	ENST00000445030	SEPTIN2-221	ENSE00003646772
2536332	ENSG00000168385	SEPTIN2	ENST00000445030	SEPTIN2-221	ENSE00003785657
2536332	ENSG00000168385	SEPTIN2	ENST00000445030	SEPTIN2-221	ENSE00001555489
2536332	ENSG00000168385	SEPTIN2	ENST00000445030	SEPTIN2-221	ENSE00003633949
2536332	ENSG00000168385	SEPTIN2	ENST00000482304	SEPTIN2-238	ENSE00003616776
2536332	ENSG00000168385	SEPTIN2	ENST00000482304	SEPTIN2-238	ENSE00003521506
2536332	ENSG00000168385	SEPTIN2	ENST00000482304	SEPTIN2-238	ENSE00001510298
2536332	ENSG00000168385	SEPTIN2	ENST00000482304	SEPTIN2-238	ENSE00001924917
2536332	ENSG00000168385	SEPTIN2	ENST00000484740	SEPTIN2-241	ENSE00001839443
2536332	ENSG00000168385	SEPTIN2	ENST00000484740	SEPTIN2-241	ENSE00001954804
2536332	ENSG00000168385	SEPTIN2	ENST00000441533	SEPTIN2-219	ENSE00003679843
2536332	ENSG00000168385	SEPTIN2	ENST00000441533	SEPTIN2-219	ENSE00003646772
2536332	ENSG00000168385	SEPTIN2	ENST00000441533	SEPTIN2-219	ENSE00001459997
2536332	ENSG00000168385	SEPTIN2	ENST00000441533	SEPTIN2-219	ENSE00001611498
2536332	ENSG00000168385	SEPTIN2	ENST00000441533	SEPTIN2-219	ENSE00001750572
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00003679843
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00003646772
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00001459997
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00001559929
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00001654879
2536332	ENSG00000168385	SEPTIN2	ENST00000420786	SEPTIN2-210	ENSE00001679913
2536508	ENSG00000168385	SEPTIN2	ENST00000481500	SEPTIN2-237	ENSE00003667943
2536508	ENSG00000168385	SEPTIN2	ENST00000481500	SEPTIN2-237	ENSE00001838749
2536508	ENSG00000168385	SEPTIN2	ENST00000481500	SEPTIN2-237	ENSE00003690740
2536508	ENSG00000168385	SEPTIN2	ENST00000481500	SEPTIN2-237	ENSE00001871675

A protein that is a component of the 26S proteosome complete is the gene Proteasome 20S Subunit Beta 2 (PSMB2), which is located in feature "2406361." This gene is linked to fibrosis and codes for a protein. It is a “non-catalytic component of the 20S core proteasome complex involved in the proteolytic degradation of most intracellular proteins.”. [14]

The protein-coding gene SEPTIN2, which is present in features "2536332" and "236508," is a "filament-forming cytoskeletal GTPase." Meningioma of the pineal area, granulomatous disease, chronic, autosomal recessive, 2, hemophagocytic lym-

phohistiocytosis, familial, 1, polycystic kidney disease 4 with/without polycystic liver disease, and Joubert syndrome 1 are all linked to the gene. [15]

Located in feature "3096512," Ribonucleic Acid (RNA), U1 Small Nuclear 124, Pseudogene (RNU1-124P) is a pseudogene that performs the function of "pre-mRNA 5'-splice site binding." [16] This gene is not linked to any illnesses.

The protein-coding gene Anaphase Promoting Complex Subunit 5 (ANAPC5), which is present in feature "3475038," "controls progression through mitosis and the G1 phase of the cell cycle." [17] Atypical hemolytic-uremic syndrome is linked to it. [17]

The protein-coding gene Asparaginyl-TRNA Synthetase 1 (NARS1), which is present in feature "3809690," is essential for the development of the cerebral cortex and for the signaling of cell migration that expresses CCR3. [18] Neurodevelopmental problems characterized by microcephaly, poor language, epilepsy, and aberrant gait are known to be caused by the mutation of NARS1. It is also associated with nuclear type 35 mitochondrial complex I deficiency and tic disorder. [18]

4 Discussion and Conclusion

For biomedical researchers, machine learning is a useful tool, particularly when it comes to identifying biomarkers. Numerous professionals in the fields of drug research, diagnostics, prognosis, and health surveillance were guided by these methodologies. It was discovered that using ML to extract biomarkers from DLPFC gene expressions was beneficial. Different dataset types were used with different machine learning approaches. While RF performed well on the multinomial dataset, SVM performed better on the binomial dataset.

The researcher began with 22,011 features in this study and, following model fitting, reduced it to 46 features for the multinomial dataset and 23 features for the binomial dataset; 10 of these features were shared by the researchers and were titled "2831519", "3302740", "3354095", "2804221", "3843934", "2406361", "2406361", "3354095", "3337168", and "2674432". None of these features had annotations.

For the multinomial dataset, 30 of the 45 features are found significant across all the classifications: "2844453", "3808600", "2674432", "3890913", "2359439", "3780271", "3475038", "2568968", "3407583", "2531129", "3556274", "3352040", "2773545", "3104698", "3556202", "3744589", "3708874", "2601230", "3061438", "3134828", "3632862", "3428268", "3048778", "3556110", "3354095", "3337168", "3556094", "2968401", and "3005266". 16 features—"3809690," "2536332", "3157311", "3556158", "3818376", "2385696", "3500787", "3405587", "2406361", "3336422", "3103062", "3346453", "3551851", "3585272", "3017068", and "2968349"—are important in both BD and MDD but not for SCZ. The features "2674317" and "3551851" on the same page are not deemed important for CTL. This

implies that all three illnesses can be predicted by the expression of a gene linked to characteristic "2674317." The feature "3157311" is the strongest predictor for all but SCZ in Table 5. To increase feature extraction and prediction capability, the researcher recommends studying schizophrenia either by itself or in conjunction with other illnesses that fall under the same DSM-V categorization.

Table 5. Multinomial's features per classification (arranged in descending order of importance)

Rank	BD	MDD	SCZ	CTL
1	3157311	3157311	2844453	3157311
2	3352040	3556094	3808600	3352040
3	3556094	3134828	2674432	3556094
4	3134828	3352040	3890913	3744589
5	3744589	3744589	2359439	3354095
6	2385696	3337168	3780271	3337168
7	3809690	3890913	3475038	2385696
8	3337168	3475038	2674317	3005266
9	2844453	3780271	2568968	2674432
10	3005266	2844453	3407583	2601230
11	2968401	2385696	2531129	3346453
12	3405587	2968401	3556274	3428268
13	2568968	3632862	3352040	3556110
14	2674432	2406361	2773545	3585272
15	2406361	3809690	3104698	3134828
16	3500787	2531129	3556202	3780271
17	3890913	3585272	3744589	2536332
18	3780271	3346453	3708874	3500787
19	3354095	2601230	2601230	3103062
20	3818376	3005266	3061438	3632862
21	3632862	3500787	3134828	3405587
22	2773545	3103062	3632862	3818376
23	3475038	3405587	3428268	3890913
24	3103062	2568968	3048778	3808600
25	2601230	3428268	3556110	3048778
26	3061438	3354095	3354095	3556158
27	3808600	3708874	3337168	2568968
28	2674317	2674317	3556094	3017068

29	3556110	2773545	2968401	2968349
30	3346453	3336422	3005266	2359439
31	3428268	3556202		2406361
32	2531129	3808600		3475038
33	3336422	3104698		2531129
34	3104698	3556158		3708874
35	3048778	3061438		3061438
36	3708874	3556274		3104698
37	2968349	2968349		3809690
38	3585272	3551851		2968401
39	3556274	3556110		3407583
40	3556158	3818376		2773545
41	3407583	2359439		2844453
42	2536332	2674432		3556202
43	2359439	3048778		3336422
44	3556202	2536332		3556274
45	3551851	3407583		
46	3017068	3017068		

Only five of the characteristics covered have Ensembl Biomart annotations available. These are NARS1 ("3809690"), RNU1-124P ("3096512"), ANAPC5 ("3475038"), SEPTIN2 ("2536332" and "236508"), and PSMB2 ("2406361"). In BD and MDD, SEPTIN2 is important, but not in SCZ. Asparaginyl-tRNA Aminoacylation, Positive Regulation of Mitotic Metaphase/anaphase Transition, Cell Division, Meiotic Cell Cycle Regulation, Cytoskeleton-dependent Cytokinesis, Anaphase-promoting Complex-dependent Catabolic Process, Reaction to Organonitrogen Compound, Regulation of Exocytosis, Cell Cycle, and Protein K11-linked Ubiquitination are the biological processes associated with these genes. Smoothened Signaling Pathway, Protein K48-linked Ubiquitination, Response to Organic Cyclic Compound, Protein Localization, MRNA 5'-splice Site Recognition, Regulation of Mitotic Cell Cycle, Proteasomal Protein Catabolic Process, Proteolysis Associated with Protein Catabolic Process, and TRNA Aminoacylation for Protein Translation.

NARS1 and SEPTIN2 were shown to be abundant genes using the Gene Card website's Gene Analytics. No matter how much transcript or protein a gene produces, its expression levels in a specimen are far higher than those of genes that are abundant. An abundant gene can support essential biological functions such as executive control, memory, cognition, and decision-making. Additionally, they may contribute to neuronal network dynamics, synaptic plasticity, and neurotransmission. However, the presence of several genes could also indicate DLPFC malfunction, which could result in mood swings and attentional changes. MDD, BD, and SCZ may be connect-

ed to the DLPFC dysfunction brought on by the overabundance of NARS1 and SEPTIN2.

Gene Analytics identified PSMB2, SEPTIN2, ANAPC5, and NARS1 as housekeeping genes. Cellular maintenance is accomplished by housekeeping genes, also known as reference genes. All cell types and tissues exhibit their expression and activity. They are essential for DNA replication, metabolism, and the control of the cell cycle from conception to death. They are frequently employed as controls or benchmarks to look for variation because of their capacity to keep their expressions stable and consistent. However, there may be some variations in housekeeping genes as well, which could result in anomalies. The four genes may be found in each patient's DLPFC in the case of DLPFC. Its presence or whether those genes changed in any way require more investigation.

On both binomial and multinomial datasets, several genes were present. The loss of proteins necessary for interphase microtubule structure from the centrosome was connected to the probes "2406361" and "2536332" for the genes SEPTIN2 and PSMB2. The concept that some people are more genetically predisposed to suffer from such disorders was further supported by the observation that NARS1 and PSMB2 directly affected a person's neurodevelopment by examining the gene functions. NARS1 may impair the cerebral cortex's ability to function normally, whereas PSMB2 in particular has a protein breakdown effect that causes neuropathy. The pathways of sister chromatid separation, cell cycle checkpoints, and pre-replicative complex assembly were associated with ANAPC5 and PSMB2.

Only NARS1 is linked to Abnormal Arachnoid Mater Morphology, Abnormal Subarachnoid Space Morphology, Aplasia/Hypoplasia of the Cerebral White Matter, Widened Cerebral Subarachnoid Space, Arachnoid Cyst, Bilateral Tonic-clonic Seizures with Generalized Onset, and Pachygyria in the section on phenotypes, specifically in the nervous system. The cerebrum is affected by these anomalies. Our senses, language, working memory, behavior and personality, movement, and cognitive processes like learning, reasoning, and logic are all controlled by the cerebrum, the biggest region of the brain. Attention-deficit hyperactivity disorder (ADHD), anxiety disorders, depression (including MDD), and post-traumatic stress disorder (PTSD) are all known to be brought on by alterations in the cerebrum. The list did not include BD. Gene Cards lacked any noteworthy data regarding the gene RNU1-124P.

Bristot's (2020) analysis did not include all five of the genes found in this investigation. It is not possible to verify in this work whether the prediction models, which use the same genes independent of the input datasets, are stable. [11] Several references or a "gold standard," which includes a list of established genes for BD, MDD, and SCZ, are necessary for stability verification. [11] Other researchers in this field of mental health may find it useful to assess the accuracy and performance of their findings in a future review paper that includes these genes.

References

[1] D. Arias, S. Saxena, S. Ephane Verguet, Quantifying the global burden of mental disorders and their economic value, (2022). https://doi.org/10.1016/j.

[2] World Health Organization, Mental disorders, (2022). https://www.who.int/news-room/fact-sheets/detail/mental-disorders (accessed April 23, 2023).

[3] V.E. Sturm, C.M. Haase, R.W. Levenson, Emotional Dysfunction in Psychopathology and Neuropathology: Neural and Genetic Pathways, in: Genomics, Circuits, and Pathways in Clinical Neuropsychiatry, Elsevier Inc., 2016: pp. 345–364. https://doi.org/10.1016/B978-0-12-800105-9.00022-6.

[4] A.L. Samy, Z.F. Khalaf, W.-Y. Low, Z. Fazli Khalaf, Mental Health in the Asia-Pacific Region: An Overview, 2015.

[5] J.W. Gardner, J.S. Sanborn, Years of potential life lost (YPLL)--what does it measure?, (1990). https://doi.org/10.1097/00001648-199007000-00012.

[6] G. Bristot, M.A. De Bastiani, B. Pfaffenseller, F. Kapczinski, M. Kauer-Sant'Anna, Gene Regulatory Network of Dorsolateral Prefrontal Cortex: a Master Regulator Analysis of Major Psychiatric Disorders, Mol Neurobiol 57 (2020) 1305–1316. https://doi.org/10.1007/s12035-019-01815-2.

[7] I. Grande, M. Berk, B. Birmaher, E. Vieta, Bipolar disorder, The Lancet 387 (2016) 1561–1572. https://doi.org/10.1016/S0140-6736(15)00241-X.

[8] J.K. Rilling, Social Interaction Decision Making and Social Interactions, 2009.

[9] V. Nejati, R. Majdi, M.A. Salehinejad, M.A. Nitsche, The role of dorsolateral and ventromedial prefrontal cortex in the processing of emotional dimensions, Scientific Reports | 11 (2021) 1971. https://doi.org/10.1038/s41598-021-81454-7.

[10] Affymetrix CEL files - Easy Guides - Wiki - STHDA, (n.d.). http://www.sthda.com/english/wiki/affymetrix-cel-files (accessed May 15, 2023).

[11] D.W.. Hosmer, Stanley. Lemeshow, R.X.. Sturdivant, Applied Logistic Regression Third edition, Wiley, 2013.

[12] X. Zhang, I. Jonassen, A. Goksøyr, Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data, Bioinformatics (2021). https://doi.org/10.36255/exonpublications.bioinformatics.2021.

[13] J.M.A. Miñoza, J.A. Rico, P.R.F. Zamora, M. Bacolod, R. Laubenbacher, G.G. Dumancas, R. de Castro, Biomarker Discovery for Meta-Classification of Melanoma Metastatic Progression Using Transfer Learning, Genes (Basel) 13 (2022). https://doi.org/10.3390/genes13122303.

[14] PSMB2 Gene - GeneCards | PSB2 Protein | PSB2 Antibody, (n.d.). https://www.genecards.org/cgi-bin/carddisp.pl?gene=PSMB2 (accessed July 19, 2024).

[15] SEPTIN2 Gene - GeneCards | SEPT2 Protein | SEPT2 Antibody, (n.d.). https://www.genecards.org/cgi-bin/carddisp.pl?gene=SEPTIN2 (accessed July 19, 2024).

[16] RNU1-124P Gene - GeneCards | RNU1-124P Pseudogene, (n.d.). https://www.genecards.org/cgi-bin/carddisp.pl?gene=RNU1-124P (accessed July 19, 2024).
[17] ANAPC5 Gene - GeneCards | APC5 Protein | APC5 Antibody, (n.d.). https://www.genecards.org/cgi-bin/carddisp.pl?gene=ANAPC5 (accessed July 19, 2024).
[18] NARS1 Gene - GeneCards | SYNC Protein | SYNC Antibody, (n.d.). https://www.genecards.org/cgi-bin/carddisp.pl?gene=NARS1 (accessed July 19, 2024).