Medical Journal of Tabriz University of Medical Sciences

Mohseni, et al. *Med J Tabriz Uni Med Sciences.* 2025;47(2):.
doi: 10.34172/mj.025.33646
https://mj.tbzmed.ac.ir

TUOMS PRESS

Original Article

Diagnosis of bipolar disorder based on extracted significant biomarkers using bioinformatics and machine learning algorithms

Hamid Mohseni[1], Massoud Sokouti[2*], Akram Nezhadi[3], Ali Sayadi[4]

[1]Department of Computer Engineering, Faculty of Electrical and Computer Engineering, Khatam Al Anbia Air Defense University, Tehran, Iran
[2]Nuclear Medicine Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[3]Cognitive and Behavioral Science Research Center, AJA University of Medical Sciences, Tehran, Iran
[4]Department of Counseling and Educational Sciences, Guilan University, Rasht, Iran

ARTICLE INFO

Article History:
Received: 27/Jan/2024
Accepted:29/Jun/2024
ePublished: 25/Jan/2025

Keywords:
- Bipolar disorder
- Biomarker
- Machine learning algorithm
- Artificial neural network
- Tree algorithm

Abstract

Background. Bipolar disorder is a type of psychiatric disease characterized by periodic mood swings that include periods of depression and mania. Military service has a significant effect on the recurrence and exacerbation of its symptoms in men with bipolar disorder. Exemption of people with bipolar disorder is one of the problems of the military service system.

Methods. First, three datasets related to bipolar disorder, including GSE53987, GSE35977, and GSE12679, were extracted from the PubMed database, which included 218 human samples and 9888458 genes. Then, genes directly related to bipolar disorder were extracted using R programming language. The shared genes were obtained from the database and extracted for 12 states with Cytoscape 3.7.1. The obtained gene expression data were trained by artificial neural network and decision tree method to identify the best models. Four parameters of sensitivity, specificity, accuracy, and area under the curve (AUC) were used to check the optimality of the model resulting from the training of machine learning algorithms.

Results. After R language preprocessing, 201 common genes were obtained. Then, 12 modes of 20 genes and 10 genes were extracted using the Cytohubba plugin in Cytoscape 3.7.1. The best model of 20 genes in the artificial neural network showed an AUC of 72% and the best model of 10 genes in the decision tree model showed an AUC of 78%.

Conclusion. We presented two models to diagnose bipolar disorder. One model was developed using artificial neural network and tanh functions and the other model was developed using decision tree algorithm.

Practical Implications. The model developed by artificial neural network and the decision tree can be used in the diagnosis of bipolar disorder in order to screen conscripts who have this disorder with a high risk of relapse and exacerbation of symptoms.

How to cite this article: Mohseni H, Sokouti M, Nezhadi A, Sayadi A. Diagnosis of bipolar disorder based on extracted significant biomarkers using bioinformatics and machine learning algorithms. *Med J Tabriz Uni Med Sciences.* 2025;47(2):. doi:10.34172/mj.025.33646. Persian.

*Corresponding author; Email: m_sokouti@yahoo.com

Extended Abstract

Background

Bipolar disorder is a mental disorder that may last for days or weeks. Bipolar people may have severe depression and suicidal thoughts. It accounts for 1.5% of all deaths in the world. In 1995, the suicide rate in soldiers was reported to be 12.5 per 100000 people, while in the first quarter of 2022, this rate was 24.3 people per 100000, and in the first quarter of 2023, it was 26.5 people per 100000. Bipolar disorder occurs in 2% of the world's population. Over 20 years, 6% of people with bipolar disorder died by suicide. Genetic factors account for 70-90% of the risk of developing bipolar disorder. Artificial intelligence can help diagnose diseases. The present study was conducted to introduce a decision support system that uses bioinformatics and machine learning algorithms to identify genes and diagnose bipolar disorders. This system is used for screening people with bipolar disorder before being called up for the military service.

Methods

The 11th generation processor with the specifications of Intel(R) Core(TM) i7-11700, frequency of 2.5 GHz, and 64 GB of RAM was employed. For processing the genes of the database, R programming language and the WGCNA R library were used. First, three datasets related to bipolar disorder, including GSE53987, GSE35977, and GSE12679, were extracted from the PubMed database, which included 218 human samples and 9-888458 genes. All gene data were extracted to R package. Pre-processing, hierarchical clustering, adding trait data, hierarchical clustering, soft thresholding, transforming gene expression matrix to adjacency matrix, calculating (Topological overlap matrix, TOM) between genes in a network, calculating dissimilarity TOM, hierarchical clustering, tree cutting operation done to define the modules, calculating eigengene modules, hierarchical clustering of eigengenes to dissimilar modules, the automatic integration of close modules, obtaining integrated colors, merging colors, and obtaining relationship between the color module and the attribute were done. Those that had a significant level were selected. An annotation file was used to find gene symbols and store gene information. The significant common genes among target datasets were screened using the Cytohubba plugin in Cytoscape 3.7.1 and 20 common hub genes were extracted for 12 methods. These methods included MCC, MNC, DMNC, degree, EPC, bottleneck, eccentricity, closeness, radiality, betweenness, stress, and clustering coefficient. Then, its gene expressions and whether it was sick or healthy were extracted. The data were normalized and Orange 3.20.1 software was used for training by a neural network with three optimization algorithms, including limited-memory BFGS, SGD, and Adam, and structural functions, including identity, logistic, tanh, and Relu models. From 5 neurons to 80 neurons with step 5 and from 10 repetition loops to 120 repetition loops with step 10, it was performed ten times on each condition. Once again, the common genes were screened using Cytoscape 3.7.1, and 10 common genes were extracted using the 12 mentioned methods. Then, the data were trained by the decision tree model using Orange 3.20.1 software. The minimum numbers of leaves from 1 to 50 were trained with step 1, and the maximum number of subgroups from 1 to 50 was trained with step 1. Using AUC, accuracy, sensitivity, and specificity, we evaluated the entire trained dataset once again using the CV method, and the model with the highest value was selected as the best model. CV means that for different modes, train 80% of the data and test 20%, and we report the average of the modes for the above statistics.

Results

In the GSE53987 dataset, 920 genes out of 54676 genes, in GSE35977 dataset, 391 genes out of 33298 genes, and in GSE12679 dataset, 461 genes out of 54676 genes were selected due to a significant relation with traits. Then, through the following formula, shared genes were obtained, which reached 201. The best method for the artificial neural network was a cluster coefficient that included 20 genes. The

artificial neural network used the limited-memory BFGS algorithm and the logistics function, which included 20 inputs, a hidden layer with 40 neurons, and a maximum of 90 training repetitions. The training structure of the artificial neural network is feed-forward backpropagation. The maximum value for sensitivity, specificity, accuracy, and AUC is 100%, indicating that the artificial neural network correctly recognizes all the genes that have been trained. The CV evaluation results are 71.15%, 72.72%, 71.96%, and 72%, respectively. Therefore, the artificial neural network in the test part recognized 71.96% of patients, 72.72% of healthy people, and 71.15% of sick and healthy people. The best method in the decision tree algorithm was eccentricity, which included 10 genes. This model had a minimum of 12 leaves with a maximum of 13 subgroups. The results of the evaluation indicated that sensitivity, specificity, accuracy, and AUC were equal to 74.76%, 69.1%, 80.76%, and 78%, respectively, which means that the decision tree learning algorithm was able to recognize 80.76% of patients, 69.1% of healthy people, and 74.76% of healthy and sick people. In addition, as a result of CV evaluation, it was equal to 67.28%, 65.5%, 69%, and 78%, respectively, indicating that the decision tree learning algorithm in the test part was able to recognize 69% of sick people, 65.5% of healthy people, and 67.28% of healthy and ill people in general.

Conclusion

Bipolar disorder is a psychiatric disease that may be associated with suicidal thoughts or attempts during periods of relapse and exacerbation of symptoms. Bioinformatics, as a multifaceted discipline which combines genetics, computer sciences, and machine learning algorithms, has a significant role in diagnosing diseases. Using bioinformatics, we extracted 20 and 10 candidate genes for bipolar disorder, and using artificial neural network and decision tree machine learning algorithms, we presented two models for diagnosing bipolar disorder, which had an acceptable diagnostic accuracy. The presented models can be used in a system for diagnosing mental disorders to screen soldiers suffering from bipolar disorder and prevent them from joining the military service to reduce suicides.

محسنی و همکاران. *مجله پزشکی دانشگاه علوم پزشکی تبریز*. فروردین و اردیبهشت ۱۴۰۴؛۴۷(۱):
doi: 10.34172/mj.025.33646
https://mj.tbzmed.ac.ir

مقاله پژوهشی

تشخیص اختلال دوقطبی بر اساس نشانگرهای زیستی مؤثر استخراج شده با استفاده از روش‌های بیوانفورماتیک و الگوریتم‌های یادگیری ماشین

حمید محسنی[۱]، مسعود سکوتی[۲*]، اکرم نژادی[۳]، علی صیادی[۴]

[۱]گروه مهندسی کامپیوتر، دانشکده برق و کامپیوتر، دانشگاه پدافند هوایی خاتم الانبیا، تهران، ایران
[۲]مرکز تحقیقات پزشکی هسته‌ای، دانشگاه علوم پزشکی مشهد، مشهد، ایران
[۳]مرکز تحقیقات علوم شناختی و رفتاری، دانشگاه علوم پزشکی آجا، تهران، ایران
[۴]گروه مشاوره و علوم تربیتی، دانشگاه گیلان، گیلان، رشت، ایران

اطلاعات مقاله

سابقه مقاله:
دریافت: ۱۴۰۲/۹/۴
پذیرش: ۱۴۰۳/۵/۱۳
انتشار برخط:

کلید واژه‌ها:
اختلال دوقطبی
نشانگر زیستی
الگوریتم یادگیری ماشین
شبکهٔ عصبی مصنوعی
الگوریتم درخت

چکیده

زمینه. اختلال دوقطبی نوعی بیماری روانپزشکی با نوسان خلقی است. مشخصهٔ این بیماری، دوره‌های افسردگی و شیدایی است.

روش کار. در پژوهش حاضر، ابتدا سه پایگاه دادهٔ GSE12679، GSE35977 و GSE53987 مرتبط با بیماری اختلال دوقطبی از پایگاه داده پابمد استخراج شد. به این ترتیب، نمونهٔ تحقیق شامل ۲۱۸ نمونه انسانی و ۹۸۸۸۴۵۸ ژن بود. سپس ژن‌هایی که با اختلال دوقطبی رابطه مستقیم داشتند، با استفاده از نرم‌افزار R استخراج شدند. اشتراک ژن‌ها از پایگاه داده‌ها به دست آمد. در نهایت، با استفاده از نرم‌افزار Cytoscape 3.7.1 ژن‌های مشترک برای ۱۲ حالت استخراج شد. برای به دست آوردن بهترین مدل‌ها، ژن‌های به دست آمده با شبکهٔ عصبی مصنوعی و درخت، آموزش داده شد. برای بررسی بهینگی از چهار پارامتر از حساسیت، تشخیص‌پذیری، صحت و ناحیهٔ زیر منحنی استفاده شد.

یافته‌ها. پس از پیش‌پردازش با نرم‌افزار R، ۲۰۱ ژن مشترک به دست آمدند. ۱۲ حالت ۲۰ ژن و ۱۰ ژن با نرم‌افزار Cytoscape 3.7.1 استخراج شدند. بهترین مدل ۲۰ ژن در شبکهٔ عصبی مصنوعی AUC برابر ۷۲ درصد و مدل ۱۰ ژن در مدل درخت AUC برابر ۷۸ درصد ارایه شدند.

نتیجه‌گیری. دو مدل برای تشخیص اختلال دوقطبی ارایه شد. یک مدل با استفاده از شبکهٔ عصبی مصنوعی و توابع logistics و مدل دیگر با استفاده از درخت بود.

پیامدهای عملی. از مدل آموزش داده شده برای غربال سربازان وظیفه مبتلا به اختلال دوقطبی با خطر عود و تشدید علایم بالا در سامانهٔ تشخیص اختلال دوقطبی می‌توان استفاده کرد.

مقدمه

اختلال دوقطبی نوعی اختلال روانی است که با دوره‌های افسردگی و دوره‌های خلق بالا مشخص می‌شود.[۱،۲] افراد دوقطبی می‌توانند احساس ناامیدی، بی‌ارزشی، گناه، از دست دادن علاقه و انگیزه، اختلال خواب و اشتها، و افسردگی شدید و افکار خودکشی داشته باشند.[۳،۴] این افراد در سیستم فعال‌سازی رفتاری، نمرهٔ بالاتری دارند و دچار اختلالات طیف رفتاری و اجتماعی می‌شوند.[۵]

تقریباً ۱/۵ درصد از کل مرگ و میرها در سراسر جهان به دلیل خودکشی است.[۶] مطالعه هلمکمپ نشان داد که در سال ۱۹۹۵ میزان خودکشی سربازان در هر صد هزار نفر، ۱۲/۵ نفر بوده است؛ این در حالی است که این میزان در یک‌چهارم اول سال ۲۰۲۲ معادل ۲۴/۳ نفر در صد هزار و در یک‌چهارم اول سال ۲۰۲۳ معادل ۲۶/۵ نفر در صد هزار بوده است.[۷-۹] بیماری‌های

*نویسنده مسؤول؛ ایمیل: m_sokouti@yahoo.com

روان‌پزشکی و اختلالات شخصیت از جمله عوامل خطر مهم در اقدام به خودکشی است.[10] نود درصد از افرادی که خودکشی کرده‌اند اختلال روانپزشکی داشته‌اند.[11] متداول‌ترین عوامل خطرآفرین برای خودکشی، بروز اختلالات روانپزشکی و اقدام به خودکشی قبلی است. اختلال افسردگی ۸۰ درصد از ۹۵ درصد اختلال‌های روانی قابل تشخیص در خودکشی کنندگان یا اقدام‌کنندگان به خودکشی را تشکیل می‌دهد.[12] از طرف دیگر، اختلال دوقطبی تقریباً در ۲ درصد از جمعیت جهان رخ می‌دهد.[13] در طی یک دوره ۲۰ ساله، ۶ درصد از مبتلایان به اختلال دوقطبی با خودکشی جان خود را از دست دادند. علل اختلالات دوقطبی به‌طور واضح شناخته نشده است اما تصور می‌شود که هر دو عامل ژنتیکی و محیطی در آن نقش دارند. پژوهشی که به بررسی افکار خودکشی سربازان نیروی زمینی سپاه انقلاب پرداخته بود نشان داد که نامناسب بودن وضعیت روانی سربازان، با افزایش افکار خودکشی در آن‌ها رابطهٔ معنی‌داری دارد.[14] وجود خودکشی میان سربازان در هر وسعت و اندازه‌ای، آثار منفی روانی و اجتماعی به دنبال دارد و برای نظام سلامت کشور و نیروهای مسلح، معضل جدی محسوب می‌شود. افرادی که به سربازی اعزام می‌شوند ممکن است از اختلالات روانی خود اطلاع نداشته باشند یا حتی آن را پنهان نمایند. در این بین، سربازان وظیفه از جمله افرادی هستند که به دلیل دوری از خانواده، جدا شدن از محیط باز و ورود به محدودیت‌های سازمانی و مقررات محدودکننده، تحت فشار روانی قرار می‌گیرند و این خود می‌تواند زمینه را برای تشدید بیماری‌های روانی فراهم کند. ملاک‌های تشخیص این اختلال از طریق مشاهدات بالینی، خودگزارشی یا اظهارات بستگان بیمار، نمرات پرسشنامه‌های مرتبط و تاریخچهٔ روانپزشکی او است. برای تشخیص این اختلال، می‌توان از آزمایش‌های خون، تصویربرداری مغزی، یا آزمون‌های روان‌سنجی برای بررسی عوامل فیزیولوژیک یا روان‌شناختی مرتبط با اختلال دوقطبی نیز استفاده کرد.[1] با این حال، به‌طور کلی روش استانداردی برای تشخیص افراد مستعد به اختلال دوقطبی در نظام‌وظیفه شناخته نشده است. از طرف دیگر، عوامل ژنتیکی حدود ۷۰ تا ۹۰ درصد از خطر ابتلا به اختلال دوقطبی را تشکیل می‌دهند.[15] اختلال ژنتیکی یک مشکل سلامتی است که در اثر یک یا چند ناهنجاری در ژنوم ایجاد می‌شود و ممکن است در اثر جهش در یک ژن (مونوژنیک) یا چندین ژن (پلی‌ژنیک) یا یک ناهنجاری کروموزومی به وجود آمده باشد.[16] برخی از بیماری‌های ژنتیکی ارثی است و برخی به دلیل جهش در ژن انجام می‌شود. تجزیه و تحلیل ژنوم انسانی با استفاده از نرم‌افزارهای رایانه‌ای از دستاوردهای بیوانفورماتیک است. یادگیری ماشین، شاخه‌ای از هوش مصنوعی است که در آن از داده‌ها و الگوریتم‌ها برای یادگیری از تجربیات استفاده می‌شود. شبکهٔ عصبی مصنوعی حداقل یک لایهٔ پنهان دارد که شامل تعداد حداقل نورون‌ها در آن است. در روش یادگیری ماشینی، شبکهٔ عصبی مصنوعی با استفاده از توابع ریاضی به‌صورت حلقه‌وار با حداقل خطای آماری آموزش داده می‌شود تا در نهایت بتوان یک مدل بهینه از آن استخراج نمود.[17] این شبکه‌ها می‌توانند در تشخیص بیماری‌ها مفید باشند.[18] لیو و همکاران، به مطالعهٔ بیماری افسردگی مفرط و ژنتیک و شبکهٔ عصبی مصنوعی پرداخته‌اند. ژو و همکاران اختلال اسکیزوفرنی و ژنتیک را با استفاده از ۵ نوع الگوریتم یادگیری ماشین مورد مطالعه قرار داده‌اند.[19،20] این ۵ الگوریتم عبارتند از: شبکهٔ عصبی مصنوعی، XGBoost extreme، تقویت‌گرادیان (Gradient boosting)، الگوریتم درخت، (Support vector machine, SVM) و (Random Forest, RF). ژانگ و همکاران با استفاده از ۵ نوع الگوریتم یادگیری ماشین، شامل یادگیری وزن‌دار محلی (Locally weighted learning)، BayesNet، بهینه‌سازی متوالی کمینه، K-نزدیک‌ترین همسایه (k-nearest neighbors)، بیز ساده (Naive Bayes) و J48، بیماری اسکیزوفرنی و ژنتیک را بررسی کرده‌اند.[21] یوجنه و همکاران بیماری دوقطبی و ژنتیک را با استفاده از الگوریتم‌های درخت و RF بررسی کرده‌اند.[22] از بین مطالعات مورد بررسی در پیشینهٔ تحقیق، این تنها پژوهشی است که به اختلال دوقطبی پرداخته است؛ اگرچه نتایج آن بر روی داده‌های آموزش داده شده تست شده و ارزیابی (Cross Validation, CV) انجام نگرفته است، از این رو، قابل اطمینان بودن مدل مورد نظر جای سؤال است. هدف از مقاله حاضر، معرفی یک سامانه تصمیم‌یار است که با استفاده از علم بیوانفورماتیک و الگوریتم یادگیری ماشین، به پردازش کد ژن‌ها و تشخیص اختلالات دوقطبی می‌پردازد. این مدل هوشمند برای نظام وظیفه طراحی شده است تا افراد را غربال کرده و از انجام خدمت سربازی منع نماید و به این ترتیب، میزان خودکشی و خودزنی در محیط نظامی کاهش یابد.

روش کار

برای پایگاه داده‌های ژنی با حجم و نمونه‌های بالا، از رایانه‌ای با پردازشگر نسل ۱۱ با مشخصات Intel(R) Core(TM) i7 -۱۱۷۰۰ و فرکانس ۲/۵ گیگاهرتز و ۶۴ گیگابایت رم استفاده شد. پردازش ژن‌های پایگاه داده با استفاده از نرم‌افزار R انجام شد. محدودیتی که بیشتر تحقیقات به آن اشاره کرده‌اند مربوط به توان

سخت‌افزاری سیستم کامپیوتر به لحاظ مقدار رم است که امکان یا عدم امکان انجام و طولانی بودن نرخ محاسباتی یا کوتاه بودن آن را میسر می‌سازد. کتابخانه‌های مورد استفاده شامل (weighted gene co-expression network analysis, WGCNA)، methylumi، survival، genefilter و readxl بودند. برای شروع کار نیاز به نمونه‌های بیان ژنی افراد بیمار و سالم بود که پایگاه داده پابمد مورد بررسی قرار گرفت تا پایگاه داده‌ای ژنی مورد نظر استخراج شود. سپس برای انتخاب ژن‌های مؤثر عملیات پیش پردازش، خوشه‌بندی سلسله مراتبی و ... که در قسمت انتخاب ژن‌های مؤثر به صورت اختصاصی آمده است انجام شد. در ادامه داده برای ورودی شبکه عصبی مصنوعی و درخت با استفاده از نرم‌افزار cytoscape 3.7.1 به حالت ۲۰ ژن و ۱۰ ژن استخراج شد. سپس آموزش داده‌ها در شبکه عصبی مصنوعی و درخت در Orange 3.20.1 انجام شد و بر اساس قسمت تحلیل داده‌ها بهترین مدل‌ها استخراج شدند. در ادامه به توضیحات بیشتر در مورد این مراحل می‌پردازیم: در ابتدا سه پایگاه داده GSE53987، GSE35977، GSE12679 مرتبط با بیماری اختلال دوقطبی از پایگاه داده پابمد استخراج شد. GSE53987 در تاریخ ۱۱ ژانویه ۲۰۱۴ رونمایی شد و منشأ آن هیپوکامپ است. این پایگاه نمونه‌های GSM1304852 تا GSM1305056 را شامل شده و حاوی ۱۰۷ نمونهٔ انسانی مرتبط است. داده‌های این پایگاه از نوع cel و حجم آن ۹۷۸/۹ مگابایت است. به عبارت دیگر، هر نمونه شامل ۵۴۶۷۶ ژن است. پایگاه GSE35977 در تاریخ ۳۱ دسامبر ۲۰۱۲ رونمایی شد. منشأ آن قشر جداری مغز است. این پایگاه شامل نمونه‌های GSM878376 تا GSM878543 می‌شود و حاوی ۹۵ نمونه مرتبط است. نوع داده‌های این پایگاه cel و حجم آن ۶۹۶/۲ مگابایت است. به عبارت دیگر، ۳۳۲۹۸ ژن در هر نمونه است. همچنین، GSE12679 در تاریخ ۱۸ دسامبر ۲۰۰۸ رونمایی شد. منشأ آن سلول‌های اندوتلیال انسانی جدا شده از قشر پیشانی پشتی جانبی پس از مرگ است. این پایگاه شامل نمونه‌های GSM318410 تا GSM318441 می‌شود که حاوی ۱۶ نمونه مرتبط است. نوع داده‌های این پایگاه cel و حجم آن ۲۲۹/۸ مگابایت است. به عبارت دیگر، ۵۴۶۷۶ ژن در هر نمونه وجود دارد. به این ترتیب، جمعاً ۲۱۸ نمونه انسانی و ۹۸۸۸۴۵۸ ژن مورد استفاده قرار گرفت. انتخاب ژن‌های مؤثر ابتدا در پایگاه داده ژنوم انسانی در R بارگذاری شد و ژن‌هایی که با اختلال دوقطبی رابطهٔ معنی‌دار داشتند، انتخاب شدند. تمام داده‌های ژنی استخراج شدند؛ به این ترتیب که هر ژن شامل چندین بیان ژن بود و به صورت آرایه ذخیره شد. در مرحلهٔ بعد پیش‌پردازش انجام شد. پیش‌پردازش شامل حذف مقادیر غایب، نرمال‌سازی داده‌ها با استفاده از روش میانه و تجزیه و تحلیل افتراقی میان نمونه‌های بیماری و گروه سالم است. حذف مقادیر غایب به این معنی است که آن دسته از نمونه ژن‌هایی که بیان ژن‌های کامل و صحیح داشتند با پرچم درست علامت‌گذاری شدند و آن دسته که به عنوان مثال حداقل یک بیان ژن نداشتند، با پرچم غلط علامت‌گذاری شدند. آن دسته از ژن‌ها که پرچم غلط داشتند، حذف شدند. نرمال‌سازی داده‌ها با استفاده از فرمول (۱) انجام شد که z_i. مقدار نرمال شدهٔ بیان ژن است.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

در مرحلهٔ بعد، بر روی داده‌های ژنی عملیات خوشه‌بندی سلسله مراتبی انجام شد. خوشه‌بندی سلسله مراتبی در علم ژنتیک به روشی اطلاق می‌شود که برای گروه‌بندی و سازمان‌دهی توالی‌های ژن و تبدیل به چند خوشه بر اساس شباهت‌هایشان به کار می‌رود. به این منظور، از اندازه‌گیری فاصله بین توالی ژن برای تعیین شباهت آن‌ها استفاده می‌شود. سپس، مجذور فاصلهٔ اقلیدسی هر نمونه با اعمال همسایگی تابع محاسبه می‌گردد. توالی‌ها به بردارهای ویژگی تبدیل شدند که دارای اطلاعاتی در مورد وقوع، مکان و رابطهٔ ترتیب k-tuples در توالی مورد نظر هستند. سپس با استفاده از یک الگوریتم خوشه‌بندی سلسله مراتبی گروه‌بندی توالی‌ها بر اساس فواصل محاسبه شده و بردارهای ویژگی انجام شد. این فرایند برای هر دنباله در خوشه خود شروع می‌شود. سپس به‌تدریج خوشه‌ها ادغام می‌شوند تا زمانی که همه دنباله‌ها در یک خوشه گروه‌بندی شوند یا سطح مطلوبی از خوشه‌بندی به دست آید. به‌طور خلاصه، خوشه‌بندی سلسله مراتبی یکی از ابزارهای قدرتمند در ژنتیک است که برای سازمان‌دهی و تجزیه و تحلیل توالی‌های ژنی به کار می‌رود. این ابزار به دانشمندان کمک می‌کند تا روابط عملکردی و الگوهای تکاملی بین ژن‌ها را کشف کنند. سپس یک آستانه‌گیری انجام می‌شود تا ژن‌هایی که به هم شباهت بیشتری دارند، انتخاب شوند. پس از انجام این مراحل، اطلاعات مربوط به صفت وارد شد تا ژن‌های بیمار و سالم را از هم تشخیص دهد. بعد، یک خوشه‌بندی سلسله مراتبی مجدد انجام شد. کل اتصال شبکهٔ نمونه، بر اساس مقیاس تابع مربع است. به همین دلیل برای به حساب آوردن شاخص‌های برازش توپولوژی که با توان‌های آستانه نرم متعدد سازگار باشند، باید از تابع برداشت آستانه نرم استفاده شود. در حوزهٔ تجزیه و تحلیل بیان ژن، آستانه نرم تکنیکی است که در ساخت شبکه‌های هم‌بیان ژن استفاده می‌شود. در این تکنیک یک توان انتخاب می‌شود که قدرت آستانه نرم نیز نامیده می‌شود و ضرایب همبستگی آن افزایش می‌یابد. این فرایند بر

همبستگی‌های قوی و کاهش تأثیر همبستگی‌های ضعیف تأکید می‌کند و تا حد قابل توجهی نویز را فیلتر می‌کند. هدف از آستانه نرم، دستیابی به یک توپولوژی بدون مقیاس در شبکه است که در آن تعداد کمی از گره‌ها (ژن‌ها) به‌عنوان هاب با اتصالات زیاد عمل می‌کنند و بیشتر گره‌ها فقط چند اتصال دارند. تصور می‌شود که به این شکل، بیولوژی زیربنایی با دقت بیشتری منعکس می‌شود، به‌طوری‌که برخی ژن‌های کلیدی بخصوص می‌توانند بسیاری از ژن‌های دیگر را تنظیم کنند. در اینجا یک رویکرد کلی برای انتخاب قدرت آستانه نرم برای یک شبکه هم‌بیان ژن وجود دارد؛ به این ترتیب که در ابتدا ضرایب همبستگی زوجی برای همه جفت‌های ژن در مجموعه داده محاسبه می‌شود. سپس، تجزیه و تحلیل توان صورت می‌گیرد؛ به این معنی که مدل توپولوژی بدون مقیاس (محور y) در برابر قدرت‌های آستانه نرم مختلف (محور x) رسم می‌شود. در نهایت توان انتخاب می‌شود. به بیان دیگر، کمترین توانی انتخاب می‌شود که شاخص تناسب توپولوژی بدون مقیاس برای آن به سطح قابل قبولی برسد. به این ترتیب، ژن‌های ویژه ماژول به‌صورت بلوکی از همه ژن‌ها محاسبه شد. سپس یک آستانه نرم اعمال شد و ماتریس بیان ژن به ماتریس مجاورت‌ها تبدیل شد. ماتریس مجاورت در واقع قدرت رابطه بین دو ژن را نشان می‌دهد. با توجه به ماتریس مجاورت، ماتریس همپوشانی توپولوژیکی محاسبه شد. ماتریس همپوشانی توپولوژیکی تشابه بین گره‌ها (ژن‌ها) در یک شبکه را نشان می‌دهد. این ماتریس، نه تنها تعامل مستقیم بین گره‌ها، بلکه اتصالات مشترک آن‌ها را نیز نشان می‌دهد. این کار با محاسبه همپوشانی توپولوژیکی برای هر جفت ژن بر اساس ماتریس مجاورت و بهبود ساختار شبکه انجام می‌شود. به این ترتیب، ماتریس همپوشانی توپولوژیکی، با در نظر گرفتن همسایگان مشترک ژن‌ها، نمایش قوی‌تری از شبکه ارایه می‌دهد. این امر می‌تواند به‌ویژه برای شناسایی ماژول‌های ژن‌های بسیار به هم پیوسته مفید باشد. خروجی این کار، ماتریسی است که در آن هر عنصر نشان‌دهندهٔ همپوشانی توپولوژیکی بین یک جفت ژن است. سپس می‌توان از این ماتریس برای تجزیه و تحلیل بیشتر، مانند خوشه‌بندی ژن‌ها در ماژول‌ها یا تجسم شبکه استفاده کرد. در ادامه، آن دسته از ژن‌هایی که با هم شباهت کمتری دارند محاسبه شدند. سپس ژن‌هایی که شباهت کم‌تری داشتند به‌صورت سلسله مراتبی خوشه‌بندی شدند. برای تعریف کردن ماژول، عملیات برش درختی انجام شد. به‌دلیل انتخاب ماژول‌های بزرگ، سایز حداقل ماژول ۳۰ در نظر گرفته شد. از برش درختی برای تجزیه و تحلیل شبکه هم‌اظهاری ژن استفاده شد. برای برش دندروگرام، از خوشه‌بندی سلسله مراتبی به ماژول‌های ژنی استفاده شد. برچسب‌های عددی ماژول‌ها به رنگ تبدیل شد. این کار برای تجسم ماژول‌ها (خوشه‌های ژن) مورد استفاده قرار می‌گیرد. در ادامه، ژن‌های ویژه ماژول محاسبه شد و برای هر ماژول شناسایی شده در شبکه هم بیان ژن محاسبه شد. ژن ویژه ماژول را می‌توان به‌عنوان نمایه بیان ژن نماینده برای یک ماژول در نظر گرفت. این الگوها، بیان ژن را در یک ماژول خلاصه می‌کند و می‌تواند برای ارتباط دادن ماژول‌ها به صفات خارجی یا یافتن محرک‌های کلیدی درون ماژول‌ها به کار برود. در مرحلهٔ بعد، ژن‌های ویژه ماژول غیرمشابه محاسبه شد. این کار امکان ارزیابی میزان تفاوت ژن‌های ویژه ماژول در مجموعه داده‌ها یا شرایط مختلف را فراهم می‌کند. این امر می‌تواند برای تجزیه و تحلیل شبکه مقایسه‌ای حیاتی باشد. سپس، خوشه‌بندی سلسله مراتبی ژن‌های ویژه ماژول غیرمشابه انجام شد. در ادامه ماژول‌های نزدیک به هم به‌صورت خودکار ادغام شدند و رنگ‌های ادغام شده به دست آمد. این کار برای ادغام ماژول‌های نزدیک و مرتبط در داده‌های بیان ژن استفاده می‌شود. شباهت بین ماژول‌ها با استفاده از ضریب همبستگی ژن‌های ویژه آن‌ها تعیین شد. به این ترتیب، ژن‌های ویژه ماژول جدید به دست آمد و به‌تبع آن رنگ‌های ادغام شده و رابطه ماژول رنگ‌ها با صفت به دست آمد و آن دسته که سطح معنی‌داری داشتند، انتخاب شدند. برای یافتن سیمبل ژن‌ها و در نهایت ذخیره اطلاعات ژن از فایل حاشیه‌نویسی استفاده شد. ژن‌های مشترک از پایگاه داده‌های اختلال دوقطبی استخراج شدند. ژن‌های مشترک در نرم‌افزار cytoscape 3.7.1 قسمت cytohubba وارد شده و از بین آنها ۲۰ ژن مشترک به ۱۲ روش ذکر شده استخراج شد. این ۱۲ روش ذکر شده شامل Neck, Bottle, EPC, Degree, DMNC, MNC, MCC, 9.5 Stress, Betweenness, Radiality, Closeness, ECCentricity, بودند. سپس بیان‌های ژن‌ها به همراه بیمار یا سالم بودن استخراج شدند. در مرحلهٔ بعد داده‌ها نرمال‌سازی شده و برای آموزش توسط شبکهٔ عصبی در نرم‌افزار orange 3.20.1 وارد شد. از سه الگوریتم Limited-memory (Broyden–Fletcher– (Stochastic gradient Descent, SGD) و Adam با مدل‌های tanh, logistic, Identity و Relu برای آموزش شبکهٔ عصبی مصنوعی استفاده شد. از ۵ نورون تا ۸۰ نورون با گام ۵ و از ۱۰ بار حلقه تکرار تا ۱۲۰ بار حلقه تکرار با گام ۱۰ آموزش داده شد و در هر یک از شرایط ۱۰ بار اجرا گردید. ژن‌های مشترک باری دیگر در نرم‌افزار cytoscape 3.7.1 قسمت cytohubba وارد شد و از بین آن‌ها ۱۰ ژن مشترک در ۱۲ روش ذکر شده استخراج گردید.

سپس بیان ژن‌ها به همراه بیمار یا سالم بودن استخراج شد. سپس داده‌ها با نرم‌افزار orange 3.20.1 توسط مدل درخت آموزش داده شد. حداقل برگ‌ها از ۱ تا ۵۰ با گام ۱ و حداکثر زیرگروه‌ها از ۱ تا ۵۰ با گام ۱ آموزش داده شد. تحلیل داده‌ها با استفاده از حساسیت، تشخیص‌پذیری، صحت و ناحیهٔ زیر منحنی یک بار تمام ست داده آموزش داده شد و بار دیگر با استفاده از روش (Cross Validation, CV) مورد ارزیابی قرار گرفت. آن‌هایی که بالاترین مقدار را داشتند در اولویت قرار گرفتند و بهترین مدل انتخاب شد. منظور از CV این است که برای حالت‌های مختلف، ۸۰ درصد از داده‌ها آموزش و ۲۰ درصد تست می‌شود و میانگین حالت‌ها برای آماره‌های بالا گزارش می‌شوند.

(۱) $Sensitivity = \frac{TP}{TP+FN}$

(۲) $Specificity = \frac{TN}{TN+FP}$

(۳) $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$

(۴) $AUC = \int_{x=0}^{1} ROC(x)dx$

منظور از TP نتیجه آزمایشی است که به درستی وجود یک وضعیت یا مشخصه را توسط الگوریتم یادگیری نشان می‌دهد و TN در واقع نتیجهٔ آزمایشی است که به درستی فقدان شرایط یا ویژگی را توسط الگوریتم یادگیری ماشین نشان می‌دهد. همچنین، FP در واقع نتیجهٔ آزمایشی را نشان می‌دهد که به اشتباه نشان می‌دهد یک شرایط یا ویژگی خاص وجود دارد و FN نتیجه آزمایشی است که به اشتباه نشان می‌دهد یک شرط یا ویژگی خاص وجود ندارد.

یافته‌ها

با بررسی پایگاه دادهٔ پابمد سه پایگاه دادهٔ GSE12679، GSE35977 و GSE83987 در ارتباط با بیماری اختلال دوقطبی به دست آمد. در پایگاه دادهٔ GSE53987، با محاسبهٔ مجدد ماژول‌های ژن ویژه از روی برچسب رنگ‌ها، رابطهٔ ماژول رنگ‌ها با صفت به دست آمده و آن دسته که سطح معنی‌دار داشتند انتخاب شدند. ضرایب همبستگی و مقادیر P در جدول ۱ ارایه شده است. به این ترتیب، ۹۲۰ ژن از ۵۴۶۷۶ ژن انتخاب شد. در پایگاه داده GSE35977، با محاسبهٔ مجدد ماژول‌های ژن ویژه از روی برچسب رنگ‌ها رابطهٔ ماژول رنگ‌ها با صفت به دست آمده و آن دسته که سطح معنی‌دار داشتند، انتخاب شدند. اطلاعات مورد نظر ضریب همبستگی و مقادیر P در جدول ۱ آمده است. به عبارتی دیگر، ۳۹۱ ژن از ۳۳۲۹۸ ژن انتخاب شد. در پایگاه داده *GSE12679*، پیش‌پردازش متد بر روی ژن‌ها انجام شد و در نهایت با محاسبهٔ مجدد ماژول‌های ژن ویژه از روی برچسب رنگ‌ها رابطهٔ ماژول رنگ‌ها با صفت به دست آمد. سپس، آن دسته که سطح معنی‌دار داشتند، انتخاب شدند. به عبارت دیگر، ۴۶۱ ژن از ۵۴۶۷۶ ژن انتخاب شد. اطلاعات مورد نظر ضریب همبستگی و مقادیر P در جدول ۱ ارایه شده است. در مرحلهٔ بعد، از طریق فرمول زیر اشتراک ژن‌ها به دست آمد که ۲۰۱ است:

(۵) $(GSE12679 \cap GSE35977) \cup (GSE12679 \cap GSE53987) \cup (GSE35977 \cap GSE53987)$

با استفاده از نرم‌افزار cytoscape 3.7.1، ۱۲ سری مدل ۲۰ بیان ژن مشترک برای پایگاه داده GSE53987 استخراج و طبق روشی که در بخش قبل توضیح داده شد، با شبکهٔ عصبی مصنوعی آموزش داده شد. بهترین روش ضریب خوشه‌ای بود که شامل ۲۰ ژن است و مشخصات آن در جدول ۲ آمده است. شبکهٔ عصبی مصنوعی از الگوریتم Limited-memory BFGS و تابع logistics استفاده کرد که شامل ۲۰ ورودی و یک لایهٔ پنهان با ۴۰ نورون و حداکثر ۹۰ بار تکرار آموزش بود. ساختار آموزش شبکهٔ عصبی مصنوعی از نوع پیش‌خور پس‌انتشار در شکل ۱ و فرمول ۶ نمایش داده شده است. با ارائهٔ مجموعه‌ای از نمونه‌های آموزشی $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ که $x_i \in R^n$ و $y_i \in \{0,1\}$ یک لایه پنهان، یک نورون پنهان MLP تابع $f(x) = w_2 g(w_1^T x + b_1) + b_2, w_1 \in R^m$ و $w_2, b_1, b_2 \in R$ را یاد می‌گیرد. w_1 w_2 وزن‌های لایه اولیه و پنهان هستند. b_1 و b_2 بایاس‌هایی که به لایه پنهان و خارجی اضافه می‌شوند را نشان می‌دهند. تابع $g(.): R \to R$ یک تابع فعال‌سازی است. تابع فعال‌سازی تابعی است که برای دریافت خروجی گره به کار می‌رود. همچنین به‌عنوان تابع انتقال شناخته شده و برای تعیین خروجی شبکهٔ عصبی، برای مثال بله یا خیر، استفاده می‌شود. مقادیر حاصله بین ۰ تا ۱ یا ۱- تا ۱ و غیره (بسته به تابع) هستند. با درنظر گرفتن تابع فعال‌ساز MLP $g(x) = \frac{1}{1+e^{-x}}, x \in R, g(x) \in [0,1]$ به صورت

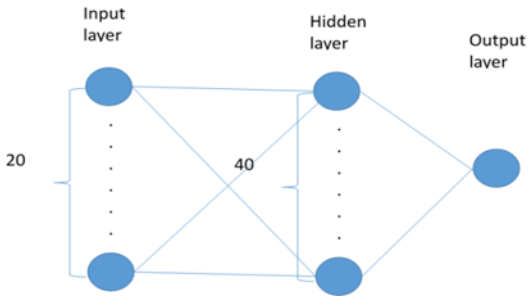
(۶) $f(x) = w2_{1\times 40} \frac{1}{1+e^{-(w1_{40\times 20} \times x_{20\times 1} + b1_{20\times 1})}}, +b_2 \ , w_1 \in R^m$ و $w_2, b_1, b_2 \in R$

بهترین نتیجهٔ به دست آمدهٔ حساسیت، تشخیص‌پذیری، صحت و AUC ، ۱۰۰ درصد بود به این معنی که شبکهٔ عصبی مصنوعی تمام ژن‌هایی که آموزش دیده را به درستی تشخیص داده است. نتایج ارزیابی (CrossValidation,CV)، به ترتیب ۷۱/۱۵، ۷۲/۷۲، ۷۱/۹۶ و ۷۲ درصد بود، به این معنی که شبکهٔ عصبی مصنوعی در قسمت تست توانسته است ۷۱/۹۶ درصد از بیماران،

۷۲/۷۲ درصد از افراد سالم و ۷۱/۱۵ درصد از افراد بیمار و سالم را تشخیص دهد. با استفاده از نرم‌افزار cytoscape 3.7.1 قسمت cytohubba ۱۲ سری مدل ۱۰ بیان ژن مشترک برای پایگاه داده GSE53987 استخراج شد و برای هرکدام با مدل درخت، طبق متد آموزش داده شد. بهترین روش ECCentricity شامل ۱۰ ژن و مشخصات آن در جدول ۲ آمده است. همچنین مدل درخت ۱۰ *ژن* بهترین مدل درخت مدلی بود که حداقل ۱۲ برگ و حداکثر ۱۳ زیرگروه داشته باشد. این مدل در شکل ۲ ارایه شده است. نتیجهٔ ارزیابی حساسیت، تشخیص‌پذیری، صحت و AUC به ترتیب برابر ۷۴/۷۶، ۶۹/۱، ۸۰/۷۶ و ۷۸ درصد بود، به این معنی که الگوریتم یادگیری درخت از بین ژن‌هایی که آموزش دیده توانسته ۸۰/۷۶ درصد از بیماران و ۶۹/۱ درصد از افراد سالم و در کل ۷۴/۷۶ درصد از افراد سالم و بیمار را تشخیص دهد. به‌علاوه، نتیجهٔ ارزیابی CV به ترتیب برابر ۶۷/۲۸، ۶۵/۵، ۶۹ و ۷۸ درصد بود به این معنی که الگوریتم یادگیری درخت در قسمت تست توانسته است ۶۹ درصد از بیماران، ۶۵/۵ درصد از افراد سالم و به‌طورکلی ۶۷/۲۸ درصد از افراد سالم و بیمار را تشخیص دهد.

جدول ۱. رنگ‌های ماژول‌ها، ضریب همبستگی و رابطه معنی‌دار آن‌ها با *P* در GSE53987، GSE35977 و GSE12679

پایگاه داده ژنی	ژن‌های منتخب	کل ژن‌ها	رنگ ماژول‌ها	ضریب همبستگی	*P*
GSE53987	۹۲۰	۵۴۶۷۶	MEmagenta	۰/۲۶۹	۰/۰۲۲
	۳۹۱	۳۳۲۹۸	MEdarkgrey	-۰/۳۰۳	۰/۰۰۳
GSE35977			MEskyblue3	-۰/۳	۰/۰۰۴
			MEtan	۰/۲۵۱	۰/۰۱۵
	۴۶۱	۵۴۶۷۶	MEsteelblue1	-۰/۷۳۲	۰/۰۰۲
			MEhotpink2	-۰/۶۶۳	۰/۰۰۷
			MEpalevioletred3	-۰/۶۴۷	۰/۰۰۹
			MEmintcream	-۰/۶۲۸	۰/۰۱۲
			MEplum1	۰/۶۰۵	۰/۰۱۷
			MEcornsilk	۰/۶۰۱	۰/۰۱۸
			MEburlywood1	-۰/۵۷۶	۰/۰۲۵
GSE12679			MEantiquewhite3	۰/۵۶۷	۰/۰۲۸
			MElavenderblush	-۰/۵۶۵	۰/۰۲۸
			MEchocolate1	-۰/۵۶۳	۰/۰۲۹
			MElightpink2	-۰/۵۳۷	۰/۰۳۹
			MEdarksalmon	-۰/۵۳۴	۰/۰۴
			MErosybrown	۰/۵۳۴	۰/۰۴
			MEcornflowerblue	۰/۵۳۲	۰/۰۴۱
			MEgoldenrod2	۰/۵۳۰	۰/۰۴۲

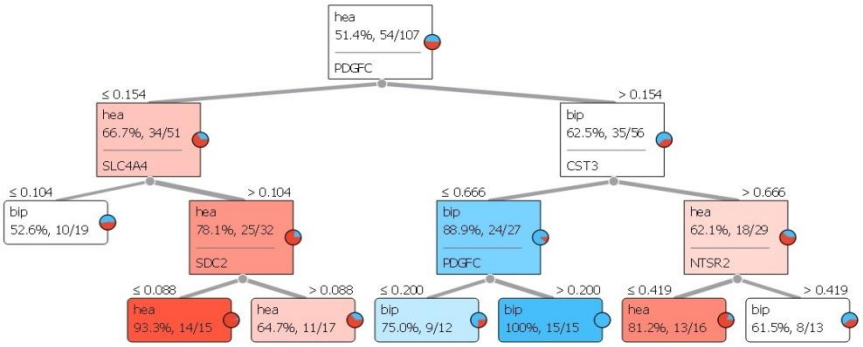


شکل۱. ساختار شبکه عصبی با ۲۰ ورودی یک لایه پنهان ۴۰ نورون و یک خروجی

جدول ۲. مشخصات ۲۰ ژن ورودی شبکه عصبی مصنوعی و ۱۰ ژن ورودی درخت

	ژن	نقش	تأثیر
روش ضریب خوشه‌ای	FGFR3	قدامی مغز، در استخوان‌سازی	کاهش آن در افسردگی مؤثر
ورودی شبکه عصبی مصنوعی	SLC7A2	سیستم L-arginine از لکوسیت‌های در گردش	در اختلال دوقطبی
	GABRG1	که کنترل اکثر سیگنال‌های مهاری در سیستم عصبی مرکزی	اختلال دوقطبی
	FXYD1	انتقال بین حوزه‌های غشایی قایق مانند در اندوزوم‌ها و غشاهای پلاسما را بر عهده دارد	اختلال دوقطبی و اسکیزوفرنی و افسردگی
	GLUD2	به‌عنوان یک هوموهگزامر برای بازیافت گلوتامات در طول انتقال عصبی	بیماری پارکینسون، اختلال دوقطبی و اسکیزوفرنی
	CYBRD1	دخالت در هومئوستاز آهن در هپاتیت C مزمن	بیماری اسکیزوفرنی و دوقطبی
	ACSS3	هومئوستاز متابولیک سلول‌های اپیتلیال را تنظیم می‌کند و فیبروز ریوی را کاهش می‌دهد	افسردگی
	ACACB	در سنتز اسیدهای چرب و مسیرهای اکسیداسیون	اختلال دوقطبی و اسکیزوفرنی
	TGFBR3	یک مهارکننده A، اما نه یک مهارکننده B، در سلول‌های گنادوتروپ در داخل بدن	اختلال دوقطبی
	DTNA	با فعال کردن STAT3 و تنظیم سیگنالینگ TGFbeta1 و P53 باعث پیشرفت کارسینوم کبدی ناشی از HBV	اختلال دوقطبی
	GPAM	آنزیم میتوکندریایی را کد می‌کند که اسیدهای چرب اشباع را به‌عنوان بستر خود برای سنتز گلیسرولیپیدها ترجیح می‌دهد.	اختلال دوقطبی
	SLCO1C1	واسطه جذب مستقل از سدیم هورمون‌های تیروئید در بافت‌های مغز است.	اختلال دوقطبی
	ALDH4A1	دستورالعمل‌هایی برای تولید آنزیم پیرولین-۵-کربوکسیلات دهیدروژناز	بیماری دوقطبی و اسکیزوفرنی
	PDGFC	در رشد، رگ‌زایی، رشد تومور، بازسازی بافت، بهبود زخم، تصلب شرایین، فیبروز، تنظیم سلول‌های بنیادی/پیش ساز و متابولیسم.نقش بسزایی دارد	افسردگی مفرط
	F3	سلول‌ها را قادر می‌سازد تا آبشارهای انعقادی خون را آغاز کنند و به‌عنوان گیرنده با میل ترکیبی بالا برای فاکتور انعقادی VII عمل می‌کند	اختلال دوقطبی
	PSD2	تنظیم انتقال سیگنال پروتئینARF و تنظیم فعالیت کاتالیزوری نقش داشته باشد	اختلال دوقطبی
	S1PR1	بازسازی قلبی ناشی از اضافه‌بار فشار را از طریق مسیر AKT-eNOS تنظیم می‌کند	اختلال دوقطبی
	METTL7A	با تنظیم متیلاسیون ژن‌های دخیل در تمایز استخوانی و بقای سلولی، بازسازی استخوان را تحریک می‌کند.	خودکشی‌های اسکیزوفرنی و دوقطبی
	SLC15A2	مسئول جذب کارنوزین به سلول‌های گلیوبلاستوما و عملکرد کامل هر سه ناقل برای حداکثر جذب مورد نیاز است	-
	GPC5	یک سرکوبگر تومور جدید که از نظر اپی‌ژنتیکی خاموش شده است، که با سرکوب سیگنالینگ Wnt/بتا-کاتنین در آدنوکارسینوم ریه، رشد تومور را مهار می‌کند	اختلال دوقطبی
ECCentricity ورودی درخت	SDC2	مهمی در پتانسیل مهاجرت سلول‌های ملانوما دارد. بیان SDC2 و CYR61 بر شدت سرطان و بقای بیماران مبتلا به کارسینوم سلول سنگفرشی مری تأثیر می‌گذارد	اختلال دوقطبی و اسکیزوفرنی
	GLUD1	اکسیداسیون برگشت‌پذیر گلوتامات به α-کتوگلوتارات و آمونیاک را کاتالیز می‌کند	اختلال دو قطبی و اسکیزوفرنی
	SDC4	برای فعال شدن اینتگرین در طول کموکینز ناشی از EGF ضروری‌است	در اختلال دوقطبی و اسکیزوفرنی

EZR	فرآیندهای بیولوژیکی[۱۰۰] از جمله چسبندگی، مهاجرت، سیتوکینز و تشکیل ساختارهای سطحی شرکت می‌کند	اختلال دوقطبی
GJA1	دستورالعمل‌هایی برای ساخت پروتئینی به نام کانکسین ۴۳ ارایه می‌دهد که یکی از ۲۱ پروتئین کانکسین است	اختلال دوقطبی
NTSR2	نقش جدید NTS در ایجاد سرطان پروستات مقاوم به اخته با تمایز نوروآندوکرین (NED) و یک استراتژی ممکن برای جلوگیری از شروع NED با هدف قرار دادن مسیر سیگنالینگ NTS	اختلال دوقطبی
SLC4A4	رمزگذاری یک انتقال‌دهنده[۱۰۳] الکتروژنی بی‌کربنات سدیم (NBCe1) که در بافت‌های طبیعی در تنظیم pH و هومئوستاز نقش دارد	اختلال دوقطبی
CST3	دستورالعمل‌هایی برای ساخت پروتئینی به نام سیستاتین C ارایه می‌دهد	-
AQP4	در توسعه، عملکرد و یکپارچگی رابط بین مغز و خون و بین مغز و مایع مغزی نخاعی نقش دارد	اختلال دو قطبی و اسکیزوفرنی
PDGFC	در رشد، رگ‌زایی، رشد تومور، بازسازی بافت، بهبود زخم، تصلب شرایین، فیبروز، تنظیم سلول‌های بنیادی/پیش ساز و متابولیسم.نقش بسزایی دارد.	افسردگی مفرط



شکل ۲. ساختار مدل درخت برای اختلال دوقطبی

بحث

در این مقاله در ابتدا ژن‌های کاندید برای بیماری اختلال دوقطبی را برای دو حالت ۱۰ ژن و ۲۰ ژن استخراج کردیم. ۲۰ ژن مهم شامل FGFR3, SLC7A2, GABRG1, FXYD1, GLUD2, GPAM, CYBRD1, ACSS3, ACACB, TGFBR3, DTNA, SLCO1C1, ALDH4A1, PDGFC, F3, PSD2, S1PR1, METTL7A, SDC2, GLUD1, SLC15A2, GPC5 بودند و ۱۰ ژن مهم شامل GJA1, NTSR2, SLC4A4, CST3, AQP4, SDC4, EZR, PDGFC بودند. در ادامه این ژن‌ها برای آموزش توسط الگوریتم یادگیری ماشین مورد استفاده قرارگرفتند. آموزش یادگیری ماشین به دو روش قابل انجام است: نوع اول با ناظر و نوع دوم بدون ناظر است. با توجه به این که داده‌های ورودی ژنی از نوع برچسب‌دار بودند، یعنی مشخص بود که کدام مربوط به افراد بیمار و کدام مربوط به افراد سالم است، از آموزش یادگیری ماشین از نوع با نظارت استفاده شد. این مدل از دقت بیشتری برخوردار است. یادگیری با نظارت یا یادگیری تحت نظارت یکی از زیرمجموعه‌های یادگیری ماشینی است. در این روش، مدل با دریافت اطلاعات برچسب خورده آموزش می‌بیند و سعی می‌کند الگوی بین داده‌ها و برچسب‌هایشان را به‌صورت یک تابع آموخته و برچسب داده‌های جدید و دیده نشده را پیش‌بینی کند. از این روش هم در مسائل طبقه‌بندی و هم در مسائل رگرسیون استفاده می‌شود.[۲۳] الگوریتم‌های یادگیری با نظارت، روند آموزشی دارند که طی آن داده‌های برچسب‌گذاری شده به الگوریتم داده می‌شود تا الگوریتم پارامترهای خودش را با استفاده از این داده‌ها

به‌روزرسانی کند و بتواند برچسب داده‌های جدید را تشخیص دهد. روش کلی به این صورت است که الگوریتم پارامترها را در جهتی تغییر می‌دهد که خطای ایجادشده در پیش‌بینی داده‌ها کمینه شود. این خطا با استفاده از تابعی که تابع هزینه نام دارد، محاسبه می‌شود.[۲۳] در این مقاله از دو الگوریتم یادگیری ماشین شبکهٔ عصبی مصنوعی و درخت برای آموزش استفاده شد. بهترین نتیجهٔ به‌دست‌آمده شبکهٔ عصبی مصنوعی پیش‌خور پس‌انتشار با ۲۰ ورودی ژن یک لایهٔ پنهان ۴۰ نورونی و یک خروجی بود. بهترین نتیجه به‌دست‌آمده حساسیت، تشخیص‌پذیری، صحت و AUC، ۱۰۰ درصد بود، به این معنی که شبکهٔ عصبی مصنوعی تمام ژن‌هایی که آموزش دیده بود را به‌درستی تشخیص داد و نتایج ارزیابی CV، به ترتیب ۷۱/۱۵، ۷۲/۷۲، ۷۱/۹۶ و ۷۲ درصد بودند، به این معنی که شبکهٔ عصبی مصنوعی در قسمت تست توانست ۷۱/۹۶ درصد از بیماران، ۷۲/۷۲ درصد از افراد سالم و ۷۱/۱۵ درصد از افراد بیمار و سالم را تشخیص دهد. به‌علاوه، نتیجهٔ ارزیابی بهترین مدل درخت حساسیت، تشخیص‌پذیری، صحت و AUC به ترتیب برابر ۷۴/۷۶، ۶۹/۱، ۸۰/۷۶ و ۷۸ درصد بود به این معنی که الگوریتم یادگیری درخت از بین ژن‌هایی که آموزش دیده بود توانست ۸۰/۷۶ درصد از بیماران و ۶۹/۱ درصد از افراد سالم و در کل ۷۴/۷۶ درصد از افراد سالم و بیمار را تشخیص دهد. همچنین، نتیجه ارزیابی CV به ترتیب برابر ۶۷/۲۸، ۶۵/۵، ۶۹ و ۷۸ درصد بود، به این معنی که الگوریتم یادگیری درخت در قسمت تست توانست ۶۹ درصد از بیماران، ۶۵/۵ درصد از افراد سالم و به‌طورکلی ۶۷/۲۸ درصد از افراد سالم و بیمار را تشخیص دهد. در تحقیق ژو و ژانگ و همکاران[۲۰،۲۱] دربارهٔ ژن‌های بیماری اسکیزوفرنی و یادگیری ماشین، تنها نتایج آموزش الگوریتم یادگیری ماشین بر روی تمام ست داده آموزش داده شده است و ارزیابی CV آن گزارش نشده است. در ژو[۲۰] شبکهٔ عصبی مصنوعی و ماشین بردار پشتیبانی مقدار AUC برای حالت ۶ ژن به ترتیب ۹۸/۹ و ۹۹/۳ درصد بوده است. همچنین، در مطالعه ژانگ[۲۱] از LWL برای آموزش ۱۰۳ ژن استفاده شده است و بهترین نتیجه AUC آن ۱۰۰ درصد بود. این در حالی است که سطح یادگیری ماشین در حالت ۲۰ ژن و۱۰ ژن دوقطبی در مدل ما ۷۸ و ۱۰۰ درصد گزارش شده است. یوجنه و همکاران[۲۲] دربارهٔ ژن‌های بیماری دوقطبی و یادگیری ماشین تحقیق کرده‌اند. در این تحقیق تنها نتایج آموزش الگوریتم یادگیری ماشین بر روی تمام ست داده آموزش داده شده است و ارزیابی CV آن گزارش نشده است. این مقاله از روش RF و الگوریتم درخت در ۶ ژن، نتیجهٔ AUC برابر ۹۶ درصد را به دست داده است درحالی‌که نتیجهٔ حالت ۲۰ ژن در مدل شبکهٔ عصبی مصنوعی ما ۱۰۰ درصد بوده است. در تحقیق لیو و همکارن[۲۴] دربارهٔ بیماری افسردگی مفرط و ژنتیک و شبکهٔ عصبی مصنوعی برای ۶ ژن AUC تنها ۷۵ درصد گزارش شده است. در هیچ یک از مقالات مورد بررسی حساسیت، تشخیص‌پذیری و صحت محاسبه نشده‌اند. از آنجا که در مدل درخت ما AUC برابر ۷۸ درصد بود ولی صحت برابر ۶۷/۲۸ درصد بود، AUC معیار خوبی برای ارزیابی نیست. زیرا در حالت شبکهٔ عصبی مصنوعی صحت مدل ما برابر ۷۱/۹۶ درصد بود ولی AUC برابر با ۷۲ درصد بود. در مقاله‌های[۲۵-۲۷] اعلام شده است که AUC بالای ۷۰ درصد معیار خوبی برای الگوریتم یادگیری ماشین است. با این حال، نتایج مقالهٔ حاضر در مدل درخت نشان داد که حساسیت، تشخیص‌پذیری و صحت کمتر از ۷۰ درصد بود درحالی‌که AUC بالای ۷۰ درصد بود. بنابراین می‌توان توصیه کرد که حساسیت، تشخیص‌پذیری و صحت هر سه بالای ۷۰ درصد باشند تا نتیجهٔ بهینگی الگوریتم یادگیری ماشین ثابت شود. در مقاله حاضر نیز از آموزش شبکهٔ عصبی مصنوعی و درخت استفاده شد. تنها در شبکهٔ عصبی مصنوعی دستیابی به این استاندارد ممکن بود.

نتیجه‌گیری

اختلال روانی دوقطبی نوعی بیماری روانی است که در دوره‌های عود و تشدید آن می‌تواند با فکر یا اقدام به خودکشی همراه باشد. علم بیوانفورماتیک که تلفیق ژنتیک و کامپیوتر و الگوریتم یادگیری ماشین است در تشخیص بیماری‌ها نقش به سزایی دارد. با این وجود، تاکنون پژوهش‌های اندکی در خصوص اختلال دوقطبی و الگوریتم یادگیری ماشین ژنتیک به انتشار رسیده است. در پژوهش حاضر، ژن‌های کاندید اختلال دوقطبی با استفاده از علم بیوانفورماتیک به صورت ۲۰ تایی و ۱۰ تایی استخراج شدند. در ادامه، با استفاده از الگوریتم‌های یادگیری ماشین شبکهٔ عصبی مصنوعی و درخت، دو مدل برای تشخیص اختلال دوقطبی از نوع ۲۰ ژن و ۱۰ ژن ارایه شد. نتایج نشان داد، این مدل‌ها دقت قابل قبولی در تشخیص بیماری اختلال دوقطبی داشتند. می‌توان از مدل‌های ارایه شده در یک سامانه تشخیص اختلالات روانی برای غربال‌گری سربازان مبتلا به اختلال دوقطبی استفاده کرد تا از اعزام آن‌ها به خدمت سربازی خودداری شود. این امر می‌تواند نقش به سزایی در جلوگیری از خودکشی سربازان وظیفه در طول دورهٔ سربازی ایفا کند. با توجه به نقش مهم سیستم‌های تصمیم‌یار و الگوریتم‌های یادگیری ماشین در حوزه‌های دیگر و تسهیل کار پزشکان از طریق ارایه سیستم‌های تصمیم‌گیری، می‌توان از این

روش در تشخیص اختلال دوقطبی نیز بهره برد. از جمله محدودیت‌های تحقیق حاضر، دسترسی به رم حداقل ۶۴ گیگابایت، پردازنده حداقل نسل ۱۰ به بالا، سرمای مناسب برای پردازشگر جهت استمرار پردازش‌ها با نرخ بالا و همچنین استفاده از هارد غیر (Solid State Disk Drive, SSD) بود. از جمله محدودیت‌های دیگر می‌توان به عدم گزارش حساسیت، تشخیص‌پذیری و صحت اشاره کرد. در نتیجه این محدودیت، در تحقیق حاضر امکان مقایسه دقیق‌تر الگوریتم‌های یادگیری ماشین فراهم نشد. نگارندگان برآنند در پژوهش آتی الگوریتم یادگیری ماشین را با الگوریتم‌های بهینه جدیدتر آموزش داده و نتایج آن را با این مقاله مقایسه نمایند. همچنین، پیشنهاد می‌شود که پایگاه داده‌های ژنومی بیشتری در پایگاه داده پابمد بارگذاری شود و همچنین یک سیستم جامع در ایران طراحی شود که داده‌های ژنومی افراد مبتلا به اختلالات روانی در آن نگهداری شود.

مشارکت پدیدآوران

حمید محسنی: ایده‌پردازی، طراحی اثر، تحلیل یا تفسیر داده‌ها، تهیه پیش‌نویس یا نقد و بررسی آن از جهت محتوای فکری؛ مسعود سکوتی: ایده‌پردازی، طراحی اثر، جمع‌آوری، تحلیل یا تفسیر داده‌ها، تهیه پیش‌نویس یا نقد و بررسی آن از جهت محتوای فکری؛ اکرم نژادی: ایده‌پردازی، طراحی اثر تهیه پیش‌نویس یا نقد و بررسی آن از جهت محتوای فکری؛ علی صیادی: نقد و بررسی آن از جهت محتوای فکری.

منابع مالی

در این پژوهش از منابع مالی مؤسسات یا سازمان‌ها استفاده نشده‌است.

دسترس‌پذیری داده‌ها

داده جدیدی تولید نشده است و از داده‌های سه لینک زیر در پابمد استفاده شده است.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12679

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35977

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse53987

ملاحظات اخلاقی

با توجه به اینکه این مطالعه هیچ داده جدیدی تولید نشده و فقط از داده‌هایی که در پایگاه پابمد به صورت عمومی در دسترس هستند استفاده شده، این مقاله ملاحظات اخلاقی را شامل نمی‌شود.

تعارض منافع

این اثر حاصل یک پژوهش مستقل بوده و هیچ‌گونه تضاد منافعی با سازمان‌ها و اشخاص دیگری ندارد.

References

1. Association AP. Diagnostic and Statistical Manual of Mental Disorders. 5 ed: Arlington, VA: American Psychiatric Pub; 2022; PP: 101-5.
2. Anderson IM, Haddad PM, Scott J. Bipolar disorder. British Mediacal journal. 2012;345:e8508. doi: 10.1136/bmj.e8508
3. Van Meter AR, Youngstrom EA. Cyclothymic disorder: a critical review. Clin Psychol Rev. 2012; 32(4): 229-43. doi:10.1016/j.cpr. 2012.02. 001
4. Bipolar Disorder. National International Mental Health; 2016`https://www.nimh.nih.gov/health/topics/bipolar-disorder'.
5. Arfaie A, Safikhanlou S, Bakhshipour Roodsari A, Farnam A, Shafiee-Kandjani AR. Assessment of Behavioral Approach and Behavioral Inhibition Systems in Mood Disorders. Basic Clin Neurosci. 2018;9(4):261-8. doi:10.32598/bcn.9.4.261
6. Fazel S, Runeson B. Suicide. New England Journal of Medicine. 2020;382:266-74.
7. Helmkamp JC. Suicides in the military:1980 – 1992. Military Medicine. 1999;160(2):45-50.
8. Military suicide rate rose 25% in first quarter, Pentagon reveals. American Military News; 2023`https://americanmilitarynews.com/2023/07/military-suicide-rate-rose-25-in-first-quarter-pentagon-reveals/'. [|updated Last Update Date|July 2023; cited Accessed Year Access Date|April 2024]|.
9. Military suicide stats released, Army saw highest increase of deaths. Military Times; 2023`https://www.militarytimes.com/news/your-military/2023/07/03/military-suicide-stats-released-army-saw-highest-increase-of-deaths/'|[updated Last Update Date|Jul 2023; cited Accessed Year Access Date| April 2024].
10. Benjamin J. Sadock, Virginia A. Sadock, Ruiz P. Kaplan and Sadock's synopsis of psychiatry behavioral science clinical psychiatry: Philadelphia: Williams and Wilkins Pub; 2007.

11. Chang B, Gitlin D, Patel R. The depressed patient and suicidal patient in the emergency department: evidence-based management and treatment strategies. Emergency Medicine Practice. 2011;13(9):1-23.
12. Delazar R, Farahi H. Suicide ideation and performing religious rites in patients with depression. Journal of Ardabil Univ Med Sci 2009;9(3):224-34.
13. Nierenberg AA, Agustini B, Kohler-Forsberg O, Cusin C, Katz D, Sylvia LG, et al. Diagnosis and Treatment of Bipolar Disorder: A Review. Journal of the American Medical Association. 2023;330(14): 1370-80. doi:10.1001/jama.2023.18588
14. Anisi J, Fathi-Ashtiani A, Soltaninejad A, M A. Prevalence of suicidal thoughts and associated factors among the soldiers. Journla Of Miliatary Medcine. 2006;2:7-11.
15. Charney A, Sklar P. Genetics of Schizophrenia and Bipolar Disorder. 5th ed: NewYork: Oxford University Press. 2018; PP:162. doi: 10.1093/med/ 9780190681425.003.0013
16. Lvovs D, Favorova OO, Favorov AV. A Polygenic Approach to the Study of Polygenic Diseases. Acta Naturae. 2012;4(3):59-71.
17. Burkov A. The hundred-page machine learning book: Andriy Burkov Quebec City, QC,Canada;2019.
18. Zolfaghari S, Sarbaz Y, Shafiee-Kandjani AR. Analysing the behaviour change of brain regions of methamphetamine abusers using electroencephalogram signals: Hope to design a decision support system. Addict Biol. 2024;29(2): e13362. doi:10.1111/adb.13362
19. Liu S, Lu T, Zhao Q, Fu B, Wang H, Li G, et al. A machine learning model for predicting patients with major depressive disorder: A study based on transcriptomic data. Frontiers in Neuroscience. 2022; 16:949609. doi: 10.3389/fnins.2022.949609.
20. Zhu L, Wu X, Xu B, Zhao Z, Yang J, Long J, et al. The machine learning algorithm for the diagnosis of schizophrenia on the basis of gene expression in peripheral blood. Neuroscience Letters. 2021;745: 135596. doi:10.1016/j.neulet.2020.135596
21. Zhang H, Xie Z, Yang Y, Zhao Y, Zhang B, Fang J. The Correlation-Base-Selection Algorithm for Diagnostic Schizophrenia Based on Blood-Based Gene Expression Signatures. BioMed Research International. 2017; 2017:1-8. doi:10.1155/ 2017/ 7860506
22. Eugene AR, Masiak J, Eugene B. Predicting lithium treatment response in bipolar patients using gender-specific gene expression biomarkers and machine learning. F1000Research. 2018;7:474. doi:10.12688/ f1000research.14451.3
23. Engelen JEv, Hoos HH. A survey on semi-supervised learning. Machine Learning. 2020;109: 73-440. doi: 10.1007/s10994-019-05855-6
24. Liu S, Lu T, Zhao Q, Fu B, Wang H, Li G, et al. A machine learning model for predicting patients with major depressive disorder: A study based on transcriptomic data. Frontiers in Neuroscience. 2022; 16(949609):1-11. doi:10.3389/fnins.2022.949609
25. Yi Z, Li Z, Yu S, Yuan C, Hong W, Wang Z, et al. Blood-based gene expression profiles models for classification of subsyndromal symptomatic depression and major depressive disorder. PLoS One. 2012;7(2):e31283. doi:10.1371/journal.pone.0031283
26. Yu JS, Xue AY, Redei EE, Bagheri N. A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder. Translational Psychiatry. 2016;6(10):e931. doi:10.1038/tp.2016.198
27. Bhak Y, Jeong HO, Cho YS, Jeon S, Cho J, Gim JA, et al. Depression and suicide risk prediction models using blood-derived multi-omics data. Transl Psychiatry. 2019;9(1):262. doi:10.1038/s41398-019-0595-2