Journal of Genetic Engineering and Biotechnology 23 (2025) 100503

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

Transcriptome Derived Artificial neural networks predict PRRC2A as a potent biomarker for epilepsy

Wayez Naqvi, Prekshi Garg, Prachi Srivastava *

Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow Campus, 226028, India

ARTICLE INFO

Keywords:
Epilepsy
PRRC2A
Artificial Neural Network
WEKA
R Studio

ABSTRACT

Epilepsy refers to the occurrence of two or more than two reiterative seizures. The occurrence of seizure is governed by the excessive electrical discharges in the cortex of the brain. Bioinformatics is crucial in diagnosing, prognosticating, and treating neurological disorders. It uses methodologies, computational tools, software, and databases to probe disease molecular underpinnings and identify biomarkers. It aids clinicians in addressing patient parameters and translational research. Artificial neural networks (ANNs) are computer models that attempt to mimic the neurons present in the human brain. This computerized neuronal model is used for analyzing and comprehending large and complex data sets. In the present study, three GEO datasets (GSE190451, GSE140393, and GSE134697) were retrieved from NCBI for the identification of differentially expressed genes using the DESeq2 package. The study identified 7 up-regulated genes (PRRC2A, FCGR3B, HLA-DRB, ENSG00000280614, ENSG00000281181, SLN, C4A) in patients with epilepsy. Furthermore, WEKA software was used for feature selection and classification of DEGs using feature selection algorithms namely Correlation Feature Selection, ReliefF, and Information Gain and classification methods such as Logistic regression, Classification via regression, Random forest, Random subspace, and Logistic model trees. After the analysis, out of the 7 genes, the C4A gene was removed as it yielded the lowest feature selection statistics. Lastly, R Studio was used for constructing the Artificial Neural Network of the 6 identified DEGs. The model's performance was evaluated using the "pROC" R package, and an AUC of 0.720 was obtained, indicating that the model had excellent classification accuracy. The NeuralNet package of R revealed that PRRC2A had the highest generalized weight value indicating the increased expression of these genes when all other parameters are constant. Therefore, PRRC2A can be used as a potential biomarker for the diagnosis of epilepsy.

1. Introduction

Epilepsy is a class of neurological disorder that is best defined as the occurrence of two or more two reiterative seizures. The occurrence of seizure is governed by the excessive and abnormal electrical discharges in the cortex of the brain[1]. A seizure can either be provoked or unprovoked[1]. Provoked seizures are the ones that occur as a result of low blood sugar levels, alcohol withdrawal, low blood sodium levels, fever, and brain infection[2]. Unprovoked seizures as the name says occur without a known cause. Stress or sleep deprivation may further aggravate this type of seizure[3]. Epilepsy affects around 50 million individuals worldwide. Based on origin, epilepsy can be categorized into three

Abbreviations: **ILAE**, International League Against Epilepsy; **TSC 1/TSC 2**, Tuberous sclerosis 1/ Tuberous sclerosis 2; **mTOR**, Mammalian Target of Rapamycin; **CSF**, Cerebrospinal Fluid; **ANN**, Artificial Neural Network; **MLP**, Multi-Layer Perceptron; **DEG**, Differentially Expressed Genes; **GEO**, Gene Expression Omnibus; **SRA**, Sequence Read Archive; **NCBI**, National Center for Biotechnology Information; **LMT**, Logistic Model Tree; **AUC**, Area Under Characteristic; **ROC**, Receiver Operating Characteristic; **C4A**, Complement C4A; **PRRC2A**, Protein-Rich Coiled Coil-2A; **PRRC2C**, Protein-Rich Coiled Coil-2C; **BAT**, Branched-chain-amino-acid Transaminase; **TNF**, Tumour Necrosis Factor; **IDDM**, Insulin-dependent Diabetes Mellitus; **TCR**, T-cell Receptor; **BCR**, B-cell Receptor; **NK cells**, Natural Killer Cells; **PI3K/AKT**, Phosphatidylinositol 3-kinase/ Protein Kinase B; **CNV**, Copy Number Variation; **HCC**, Hepatocellular Carcinoma; **CD8**, Cluster of Differentiation 8; **TGF-β**, Transforming Growth Factor-Beta; **PD-1**, Programmed Cell Death; **CTLA-4**, Cytotoxic T-Lymphocyte Antigen 4; **CD160**, Cluster of Differentiation 160; **OLIG2**, Oligodendrocyte Transcription Factor 2; **OPC**, Oligodendrocyte Progenitor; **RA**, Rheumatoid Arthritis; **PADI**, Peptidyl Arginine Deiminase; **TRAF 1**, TNF Receptor Associated Factor 1; **TFF3**, Trefoil Factor 3; **IL21**, Interleukin-21; **HLA-DQ4**, Human Leukocyte Antigen DQ4; **DESeq2**, Differential Expression Analysis based on the Negative Binomial Distribution.

* Corresponding author.
E-mail address: psrivastava@amity.edu (P. Srivastava).

https://doi.org/10.1016/j.jgeb.2025.100503
Received 18 October 2024; Received in revised form 12 March 2025; Accepted 28 April 2025
Available online 12 May 2025

different types namely: acquired/symptomatic, cryptogenic, and idiopathic. The acquired type of epilepsy is characterized by an identifiable cause. These seizures result from any sort of trauma, the presence of a tumor, or from meningeal infections. The cryptogenic type of epilepsy, as the name says is provoked by an unidentified cause. Idiopathic type of epilepsy also refers to the epilepsy provoked by an unidentifiable cause however it has a genetic basis. Epileptic seizures can be further classified broadly into two categories: Convulsive and Non-convulsive. Convulsive seizures are the most common type of seizures which involve repeated involuntary relaxation and contraction of muscles in the body thereby leading to uncontrolled shaking [4]. Non-convulsive seizures occur as a result of altered mental status[5,6]. Generalized seizures are those which affect both the hemispheres of the brain and impair consciousness with no apparent cause. Whereas focal seizures are those which affect only one hemisphere of the brain. Epilepsy is a neurological disorder, most accurately characterized by (i) the occurrence of two unprovoked seizures after a 24-hour gap, (ii) the occurrence of a single unprovoked seizure and a probability of manifestation of other seizures over 10 years, (iii) the identification of a specific epilepsy syndrome through diagnostic evaluation. These definitions and classifications are established by ILAE (International League Against Epilepsy), which aids in gaining a better understanding of epilepsy and developing strategies to eradicate it. The major factors contributing to the development of epilepsy include genetic and acquired factors however in most cases the main factor which is responsible for the development of epilepsy remains unknown[1]. Epilepsy occurs as a result of single gene defects[7]. This single gene defect occurs in genes that code for ion channels, enzymes, and G-protein coupled receptors thereby affecting the functionality of these biological entities[8]. Phakomatoses are another important factor promoting epilepsy, they are a group of diseases that affect ectodermal structures such as the central nervous system, skin, and eyes. A vast majority of phakomatoses are single-gene disorders that may be inherited in an autosomal dominant, autosomal recessive, or X-linked pattern[9]. One such example is the Tuberous sclerosis complex caused by mutations the in TSC1 or TSC2 gene which results in the up-regulation of the mTOR pathway leading to the growth of tumors in multiple organs of the body and eventually promoting an increased level of neural excitability[10,11]. GABAergic neurons which are regarded as the inhibitory neurons are lost, leading to an increased hyperexcitability of the neural networks[12]. Excessive release of the neurotransmitter glutamate after a brain injury results in excitotoxicity, which causes it to be excessively depolarized, intracellular Ca2 + concentrations sharply increase, eventually resulting in cellular damage or death[13]. Under BBB disruption, albumin was found to leak from the blood into the brain parenchyma and activate the transforming growth factor beta receptor inducing epileptogenesis[14–16]. Bioinformatics plays a key role in understanding and managing neurological disorders by helping with early diagnosis, predicting disease outcomes, and finding effective treatments. It uses various methods, computational tools, software, and databases to study the molecular causes of diseases and identify novel biomarkers. Bioinformatics helps clinicians address fundamental questions and inquiries based on every individual patient parameter, encompassing disease attributes, laboratory findings, proteomic, genomic, and metabolic data, along with other pertinent information. Bioinformatics has also helped turn research findings into real-world medical applications, supporting the discovery of new drugs and diagnostic markers. Although many potential biomarkers have been found using computer-based methods, only a few have been fully tested and confirmed in clinical trials.[17]. A major challenge with the rapid growth of bioinformatics is the huge amount of data being produced. This large volume of data makes it difficult to analyze using traditional methods, which are no longer effective. As a result, getting accurate answers has become a tough task.[18]. Machine learning, situated at the intersection of various disciplines within bioinformatics, encompasses a category of algorithms driven by data analysis. These algorithms aim to address specific issues by scrutinizing patterns within datasets, often focusing on one specific factor[10]. The application of these methodologies, renowned for their adaptability and efficacy, has gained extensive traction in the realm of biology, notably in investigations centered on the discovery of biomarkers[19,20]. This widespread utilization has given rise to a diverse array of machine learning algorithms and methodologies[21,22]. Artificial neural networks (ANNs) are computer models that attempt to mimic the neurons present in the human brain. This computerized neuronal model is used for analyzing and comprehending large and complex data sets. The learning process in Artificial Neural Networks (ANNs) is determined by how the various network components are mathematically connected. This enables the network to detect patterns in data by assigning numbers (also known as weights) to inputs and adjusting them as more data is processed, allowing the network to improve over time. ANNs' primary advantages are their ability to handle errors well and make accurate predictions or classifications, even for new or unlearned data that they have never seen before. A neural network is made up of three parts: an input layer, a hidden layer, and an output layer. The input layer, as the name suggests, contains input features. The 'hidden' layer refers to the mathematical calculations performed by the model. The output layer is the last, containing the network's output data. This makes them ideal for biomarker studies which resulted in their use in generating panels of biomarkers[23]. The architectural foundation of Artificial Neural Networks (ANNs) is rooted in the perceptron, a singular artificial processing neuron endowed with adjustable weights, a bias, and an activation threshold. However, the perceptron is limited to classifying non-linearly separable patterns, relying on error occurrence during testing for learning. In practice, ANNs typically employ a Multi-Layer Perceptron (MLP), a configuration comprising multiple perceptrons. The input data is processed in steps: first, they handle the input variables, then they use activation functions to detect important features, and finally, they generate the output or result[24]. The transcriptome denotes the complete collection of RNA transcripts within a specific cell, of a defined developmental stage or physiological state[25]. A comprehensive understanding of the transcriptome is imperative for elucidating the functional constituents of the genome and deciphering the underlying mechanisms governing development and pathological conditions[26]. High-throughput RNA-level investigations have historically employed microarray technologies, facilitating the identification of differentially expressed genes across developmental stages or between cohorts of healthy and diseased subjects[27]. However, the emergence of RNA-seq, driven by advancements in sequencing technologies, has rapidly supplanted microarray methodologies due to its superior resolution and heightened reproducibility[28,29].

In this study, comparative analyses of gene expression data were performed to identify differentially expressed genes with a prime aim to ultimately identify potent markers for epilepsy. These genes were then subjected to machine learning algorithms for feature selection and classification of DEG's. Lastly, Artificial neural networks were constructed to identify potent biomarkers[30,31].

2. Materials and methods

2.1. Data Retrieval

63 Datasets were identified from GEO[32] and SRA[33] databases of NCBI[34], as these datasets were transcriptomic. Out of the 63 identified studies 3 were selected and subsequently utilized in further research. These studies were selected because all of them affected the Temporal Neocortex region of the brain and led to the development of Temporal Lobe Epilepsy. One of the studies also showed that the PAX6 cells in the brain were being affected[35]. Fig. 1.

2.2. Pre-processing and alignment of data

FASTQC tool was used to perform the quality check on each sample. FASTQC is designed to offer an easy and efficient way to perform quality

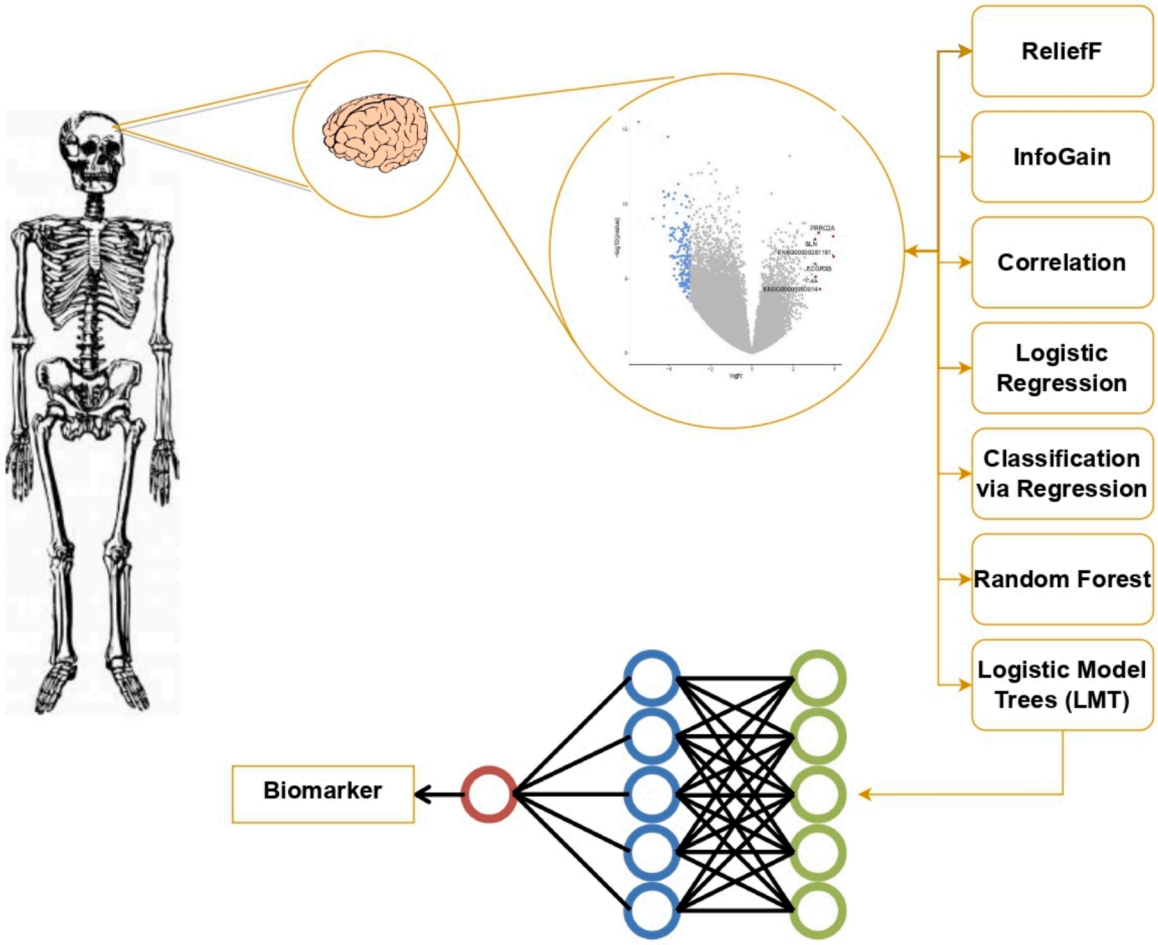

Fig. 1. Flowchart depicting the pipeline adopted in the present study.

checks on raw sequence data generated from high-throughput sequencing. It filters out low-quality sequences, resulting in extremely accurate results[38]. Following the quality check, the sequences were trimmed to remove the adapter sequences, this task was performed using the Trimmomatic tool[39]. Trimmed Samples were then aligned with a reference genome of *Homo sapiens* and were done using HISAT2. It enables extremely fast and sensitive read alignment, particularly for reads spanning two or more exons[40]. Then FeatureCounts tool was used to count RNA-Seq reads for understanding the genomic features[41].

2.3. Identification of differentially expressed genes

The differentially expressed genes in the samples were analyzed using the DESeq2 Tool[42]. This tool generates output in tabular files as well as graphical results in PDF format. To identify the up-regulating genes the tabular file containing the data about the DEGs was filtered based on the p adjusted and log2 fold chain value.

2.4. Feature selection and classification of DEGs using Machine learning algorithm

The Weka software[43] was used for feature selection. Important features were first identified using three feature selection algorithms: Information Gain[44], Correlation Feature Selection[45], and ReliefF[46]. Then, to predict genes in the up and down categories, five widely used classifiers i.e Logistic Regression[47], Classification Via Regression[48], Random Forest[49], LMT[50], and Random Subspace[51] were employed, which have been applied to solve various classification and prediction problems in biology, which showed comparable or even higher performance results than other commonly used machine learning algorithms.

2.5. Biomarker discovery using Artificial neural network

After feature selection, the artificial neural network was constructed using the NeuralNet package available in R studio[52]. RStudio is a data analysis tool for importing, accessing, transforming, plotting, and modeling data, as well as for machine learning to make predictions on data[53]. The artificial neural network constructed usually contains a single input, hidden, and output layer. To further validate the neural network, the area under characteristic (AUC) and receiver operating characteristic (ROC) curve of the training set in the "pROC" R package was used[54].

3. Results

3.1. Data collection

The Datasets were collected using NCBI, GEO, and SRA. Only those were selected which were transcriptomic, had epilepsy as the disease, and whose expression profiling was performed by High Throughput Sequencing. Out of the 63 identified studies only 3 of them were selected for further research because they had a significantly smaller number of patients and control sample ratio and affected the neocortex region of the brain Table 1.

Table 1
Details of the Three Datasets obtained from NCBI.

Accession Number	Sample Size	Condition	No of Patients	No of Control	Citations
GSE190451	6	Epilepsy	3	3	—-
GSE140393	6	Epilepsy	3	3	[36]
GSE134697	19	Epilepsy	17	2	[37]

3.2. Identification of differentially expressed genes

The final three studies were then subjected to pre-processing, following which the differentially expressed genes (DEGs) were obtained using the DESeq2 tool. All these tasks were underdone using the Galaxy server. The obtained results were in graphical and tabular form. From the list of DEGs obtained, a total of 7 genes were found to be up-regulated by setting up the p-adjusted value at less than 0.01 and log 2-fold chain value at 3Fig. 2, Table 2.

3.3. Feature selection and classification of DEGs using Machine learning algorithm

Feature selection algorithms namely Correlation, ReleifF, and Information Gain were used to improve the performance of the model by reducing the dimensionality of the data while retaining the most informative features. The F measure was calculated using these three feature selection algorithms in conjunction. After feature selection classification was done, where a class is to be assigned to the data based on its feature. Classification was done using Logistic Regression, Random Forest, Random Subspace, Logistic Model Trees, and Classification via regression. Overall, ReliefF and Correlation outperformed the InfoGain feature selection algorithm in terms of optimal Receiver Operating Characteristic (ROC) curve and Area under the ROC Curve (AUC), resulting in the highest accuracy. Out of the five classification methods used Random Subspace outperformed the other methods tested, yielding an accuracy of 80.64 % and F measure of 0.782. Thus through the following analysis, the model based on ReliefF and Correlation feature selection and Random Subspace classification was deemed the best Table 3.

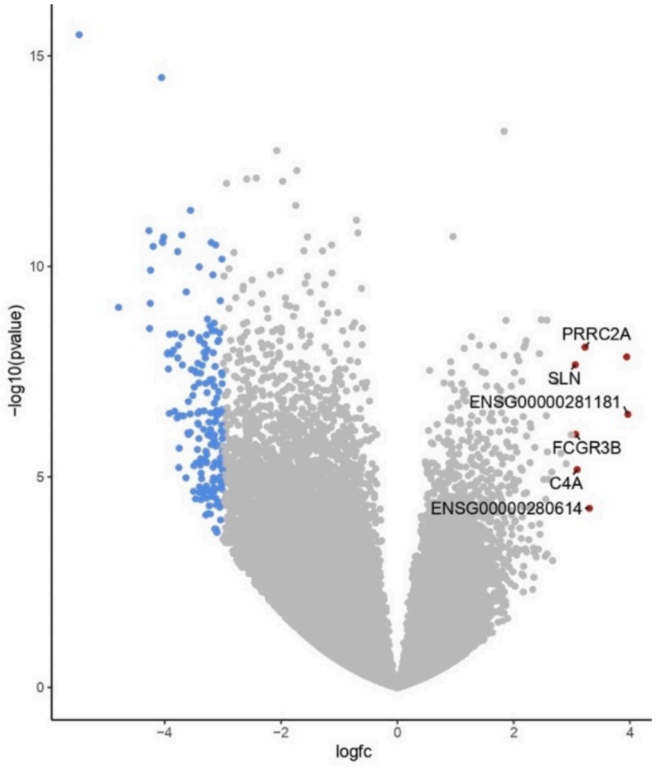

Fig. 2. Volcano plot depicting the differentially expressed genes obtained through DESeq2. Where x-axis represents the log 2 (fold chain) value and y-axis represents the −log10 (adjusted p value). Blue coloured dot represent the down-regulated genes, red coloured dot represent up-regulated genes and grey coloured dot represent the non-significant genes.

Table 2
List of the 7 up-regulating genes obtained after setting up the log 2-fold chain value at 3 and p-adjusted value at less than 0.01.

Gene ID	Gene Name	log2(FC)	P-adj
ENSG00000204469	PRRC2A	3.228347498	2.84E-06
ENSG00000198502	HLA-DRB5	3.943874933	4.00E-06
ENSG00000170290	SLN	3.061719734	5.31E-06
ENSG00000281181	Novel gene	3.965287116	3.43E-05
ENSG00000162747	FCGR3B	3.069284479	7.10E-05
ENSG00000244731	C4A	3.092539827	0.000259113
ENSG00000280614	Novel gene	3.303061069	0.001035682

3.4. Biomarker discovery using Artificial neural network

Following feature selection and classification, the artificial neural network was constructed using the pROC package, following the construction of ANN, ROC-AUC plots were made to quantify the overall performance of the model by calculating the area under the ROC curve. The higher the value of AUC the better will be the model's performance. Before constructing the ANN, out of the 7 up-regulating genes one was removed i.e. C4A as it interfered with the accuracy and f-measure value. The ANN diagnostic model had six input layers and two hidden layers. The model's performance was then evaluated and an AUC of 0.720 was obtained, indicating that the model had excellent classification accuracy. After that, the generalized weight (gw) plots were constructed using the NeuralNet package of R and analyzed, which depicted that while keeping all the attributes constant, the genes PRRC2A and HLADRB5 were found to be positively regulated while all the remaining genes were being negatively regulated Fig. 3. After analyzing the generalized weight plots, the ROC curve was constructed for further validation and lastly, out of the 6 biomarkers, only 1 was selected i.e. PRRC2A as it had the highest value Fig. 4.

4. Discussion

The RNA-Seq data analysis yielded valuable insights into potential biomarkers that could play important roles in disease diagnosis, prognosis, and therapeutic targeting. Galaxy server was used to identify differentially expressed genes associated with epilepsy. These genes appear to be promising candidates for additional research and validation as potential biomarkers. The analysis of differential gene expression assisted in identifying genes that are either being up-regulated or down-regulated. These genes provide important information on the underlying

Table 3
Values of the feature selection methods used for the classification of the seven hub genes. Out of the seven genes, one gene i.e. C4A was removed as after the feature selection algorithm it yielded the lowest values.

Gene Name	ReliefF	InfoGain	Correlation
PRRC2A	0.046	0	0.383
FCGR3B	0.034	0	0.841
HLA-DRB	0.013	0.271	0.287
ENSG00000280614 Novel Gene	0.010	0	0.256
ENSG00000281181 Novel Gene	0.010	0	0.247
SLN	0.006	0.341	0.232
C4A	−0.002	0	0.287

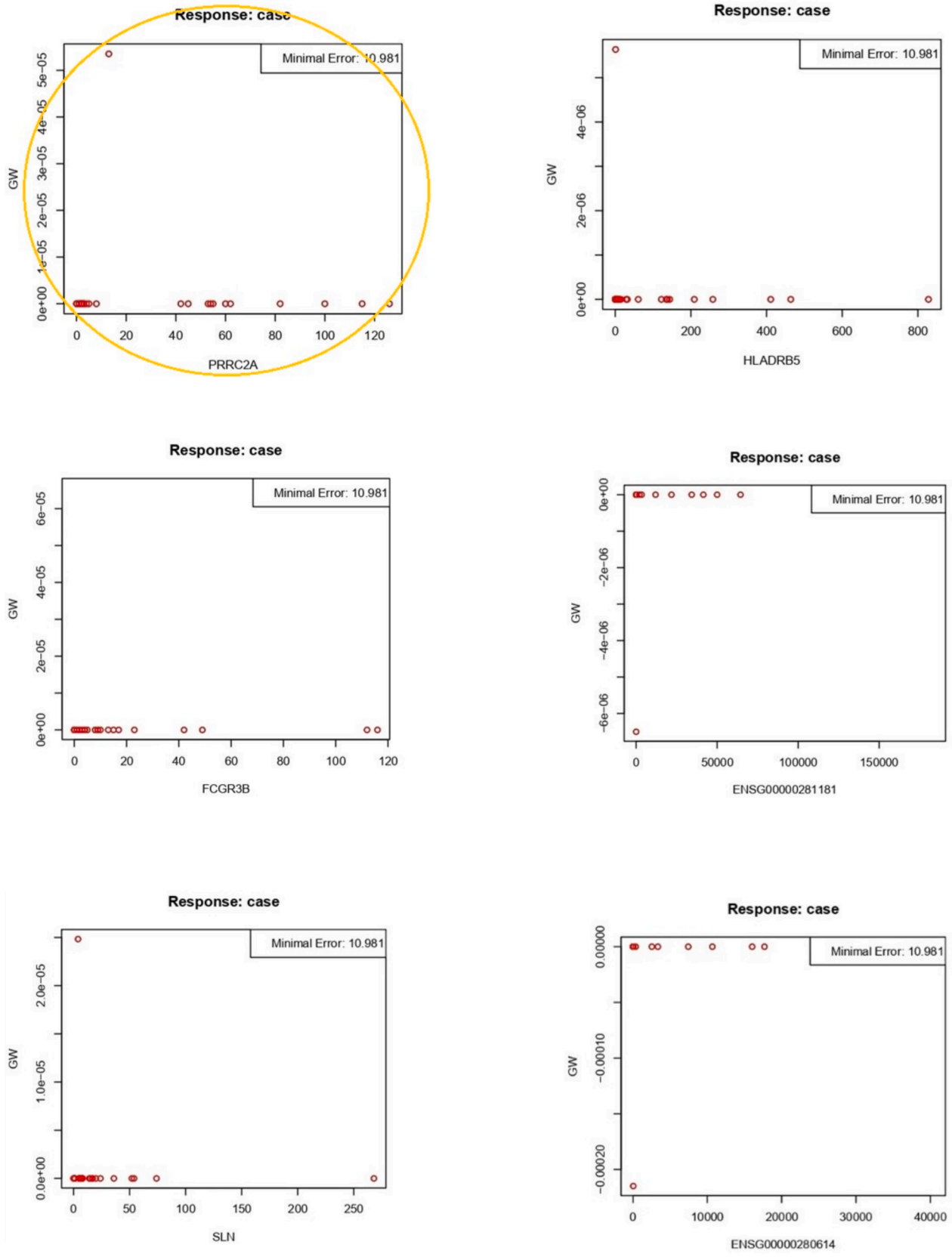

Fig. 3. Image showing response of genes demonstrated through generalized weights (GW) plot of differentially expressed genes constructed using NeuralNet package of R. The response of genes PRRC2A and HLADRB5 is seen to be positively regulated, SLN, ENSG00000281181 and ENSG00000280614 are found to be negatively regulated, whereas the response of FCGR3B is neutral.

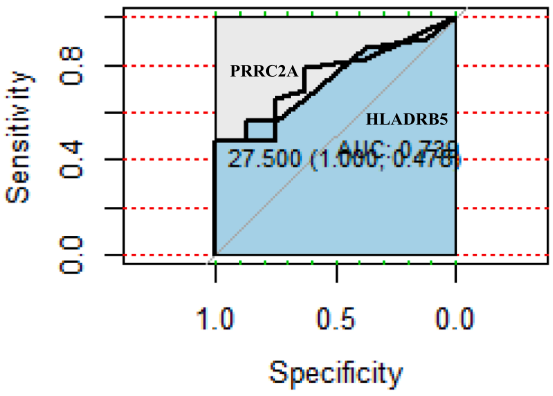

Fig. 4. ROC curve. Where X-axis represents the specificity while Y-axis represents the sensitivity.

molecular mechanisms that contribute to the disease's pathogenesis[55]. Furthermore, machine learning algorithms were used to create predictive models that can distinguish between disease and healthy samples based on the expression levels of selected genes. These models performed well, implying that these biomarkers could be used to develop diagnostic or prognostic tests. The study revealed PRRC2A as a potent biomarker in epilepsy. PRRC2A stands for Protein-rich coiled coil-2A genes. They are also known as BAT2 genes. The family of BAT genes is found close to the TNF alpha and TNF beta genes. These genes are all found to be present in the class III region of the human major histocompatibility complex. This gene has also been linked to the development of rheumatoid arthritis[56]. SNP stands for Single Nucleotide Polymorphisms and are the most common forms of genetic variations found within an individual. They contribute to genetic diversity but their role is not just restricted to this. These variations occur in either the coding or non-coding regions of the gene. These variations affect the functioning of the gene and thereby promote disease susceptibility. Hence various genotyping methods are used to map the disease-causing genes[57]. In the case of PRRC2A, certain SNP's are found to be associated with the development of rheumatoid arthritis. In one study it was identified that SNPs are linked to RNA-modified RA-susceptible genes. The SNP influences gene expression in blood cells and alters the protein levels which eventually contribute to the development of rheumatoid arthritis. For instance, SNPs were found to be associated with the expression of genes PADI2 and TRAF1 where PADI2 is involved in producing citrullinated proteins targeted by RA-specific antibodies while TRAF1 regulates the inflammation and RA disease activity. These SNP's alter the level of protein's TFF3, IL21, and HLA-DQA2. These proteins take part in the immune response and have altered levels in RA patients[58]. N6-methyladenosine (m6A) is the most common internal mRNA modification in eukaryotes. A series of enzymes, including m6A methyltransferases, demethylases, and m6A-specific binding proteins, dynamically regulate m6A modification. The discovery of new m6A-specific binding proteins in neural chelps helps in better understanding the role of 6A modification in the development of neurological disorders[59,60]. PRRC2A and PRRC2C were identified as possible m6A-binding proteins. PRRC2A was found to be more abundant in all types of neural cells than PRRC2C[61]. SNP's in the PRRC2A gene is found to be associated with multiple sclerosis[62]. PRRC2A is an m6A reader which is the methylation of binding proteins that function by down-regulating genes. In one such study, it was seen that numerous oligodendroglial-specific genes were down-regulated. One particular gene was Olig2 which stands for oligodendrocyte transcription factor 2 and controls OPC specification, differentiation, and myelination. Mouse model studies exhibited that knockout of PRRC2A resulted in hypomyelination and cognitive defects[63] . SNP's of PRRC2A are found to be associated with various types of cancers such as breast cancer, lung cancer, hepatocellular carcinoma, and non-Hodgkin lymphoma[64]. High levels of PRRC2A are characterized by CNVs and DNA methylation[65]. High levels of PRRC2A correlate with a high alpha-fetoprotein level and poor differentiation grade[64]. The variant of PRRC2A is involved in the cell proliferation and TGF-β signaling pathway. Knockdown of PRRC2A inhibited the proliferation, migration, and invasion of HCC cells[64]. The tumor immune microenvironment plays a major role in the progression and metastasis of cancer including HCC[66]. CD 8 T-cell destroys the tumor cells through the action of perforins and granzymes[67]. Studies reveal that PRRC2A strengthens the microenvironment of HCC cells by barricading them from the entry of CD8 T cells[68]. Tumour cells express ligands that can bind with corresponding proteins on the T cell which prevents the secretion of cytokines. This eventually leads to T-cell exhaustion. PD-1, CTLA-4, and CD160 are some of the receptors affected by this condition[64,69]. PRRC2A is also found to be correlated with Type 1 Diabetes. PRRC2A gene contains microsatellite repeats and missense polymorphisms[70]. These variations are found to be linked with the development of Type 1 Diabetes. High levels of PRR2CA are linked with a significantly enriched PI3K/AKT signalling pathway. This pathway regulates the immunity of an individual. This pathway is initiated by either a BCR or TCR and regulates lymphocyte differentiation. High levels of PRRC2A are correlated with enhanced functionality of pathogenic immune cells like effector memory T cells, NK cells, and Plasma cells[71]. Neurotransmitters like acetylcholine and catecholamines modulate the immune system. They do so by binding to receptors present in monocytes and lymphocytes. Both of these phenomena lead to an imbalanced immune homeostasis in Type 1 Diabetes[70,72].

5. Conclusion

Epilepsy is a serious neurological disorder affecting millions worldwide. This study aimed to identify differentially expressed genes (DEGs) in epilepsy patients, where a gene is considered differentially expressed if its expression level changes under specific conditions. Using the DESeq2 pipeline, datasets from GEO were analyzed, revealing seven up-regulated genes. These genes then underwent feature selection before being processed through artificial neural networks. Generalized weight (GW) graphs were generated, identifying PRRC2A and HLADRB5 as positively regulated, with PRRC2A showing the highest value. These findings suggest that PRRC2A could serve as a potential biomarker for epilepsy. However, while these results are promising, extensive experimental validation is necessary. Given that machine learning models are prone to biases, rigorous testing is essential to confirm the functional role of these biomarkers.

CRediT authorship contribution statement

Wayez Naqvi: Writing – review & editing, Writing – original draft, Investigation, Data curation. **Prekshi Garg:** Methodology, Conceptualization. **Prachi Srivastava:** Supervision.

Funding

None.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the bioinformatics tools and software used in the present study. The authors are grateful to Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow

campus where the benchwork for the present study was done.

References

1. Fisher RS, Acevedo C. ILAE official report: a practical clinical definition of epilepsy. *Epilepsia*. 2014;55(4):475–482. https://doi.org/10.1111/epi.12550.
2. Murray EL, Kansagra A. *Essentials of Hospital Neurology*. Oxford University Press; 2017.
3. Wilden JA, Cohen-Gadol AA. Evaluation of first nonfebrile seizures. *Am Fam Physician.* 2012;86(4):334–340.
4. National Clinical Centre. The Epilepsies: The diagnosis and management of the epilepsies in adults and children in primary and secondary care. National Institute for Health and Clinical Excellence; 2013 Dec 16.
5. Hughes JR. Absence seizures: a review of recent reports with new concepts. *Epilepsy Behav*. 2009;15(4):404–412. https://doi.org/10.1016/j.yebeh.2009.06.007.
6. Chang AK, Shinnar S. Nonconvulsive status epilepticus. *Emerg Med Clin North Am.* 2011;29(1):65–72. https://doi.org/10.1016/j.emc.2010.08.006.
7. Pandolfo M. Genetics of epilepsy. *Semin Neurol.* 2011;31(5):506–518. https://doi.org/10.1055/s-0031-1299789.
8. Berkovic SF, Howell RA, Scheffer IE. Human epilepsies: interaction of genetic and acquired factors. *Trends Neurosci.* 2006;29(7):391–397.
9. Stafstrom CE, Sutula TP. Epilepsy mechanisms in neurocutaneous disorders: tuberous sclerosis complex, neurofibromatosis type 1, and Sturge–Weber syndrome. *Front Neurol.* 2017;17(8):87. https://doi.org/10.3389/fneur.2017.00087.
10. Northrup H, Krueger DA. Updated international tuberous sclerosis complex diagnostic criteria and surveillance and management recommendations. *Pediatr Neurol.* 2021;123:50–66. https://doi.org/10.1016/j.pediatrneurol.2021.07.011.
11. Curatolo P. Mechanistic target of rapamycin (mTOR) in tuberous sclerosis complex-associated epilepsy. *Pediatr Neurol.* 2015;52(3):281–289. https://doi.org/10.1016/j.pediatrneurol.2014.10.028.
12. Armijo JA, Valdizán EM. Advances in the physiopathology of epileptogenesis: molecular aspects. *Rev Neurol.* 2002;34(5):409–429.
13. Aroniadou-Anderjaska V, Fritsch B, Qashu F, Braga MF. Pathology and pathophysiology of the amygdala in epileptogenesis and epilepsy. *Epilepsy Res.* 2009;85(2–3):102–116. https://doi.org/10.1016/j.eplepsyres.2007.11.011.
14. Ivens S, Kaufer D, Flores LP. TGF-β receptor-mediated albumin uptake into astrocytes is involved in neocortical epileptogenesis. *Brain*. 2006;129(5):535–547. https://doi.org/10.1093/brain/awl317.
15. Cacheaux LP, Miglioretti M, Kwon T. Transcriptome profiling reveals TGF-β signaling involvement in epileptogenesis. *J Neurosci.* 2009;29(28):8927–8935. https://doi.org/10.1523/JNEUROSCI.0430-09.2009.
16. David Y, Cacheaux LP. Astrocytic dysfunction in epileptogenesis: consequence of altered potassium and glutamate homeostasis? *J Neurosci.* 2009;29(34):10588–10599. https://doi.org/10.1523/JNEUROSCI.2323-09.2009.
17. Suh KS, Schrader SS. Tissue banking, bioinformatics, and electronic medical records: the front-end requirements for personalized medicine. *J Oncol.* 2013;2013:1–12. https://doi.org/10.1155/2013/368751.
18. Singh S, Kumar S. Biomarkers for detection, prognosis and therapeutic assessment of neurological disorders. *Rev Neurosci.* 2018;29(8):771–789. https://doi.org/10.1515/revneuro-2017-0097.
19. Cook CE, Taylor M. The European bioinformatics institute in 2016: data growth and integration. *Nucleic Acids Res.* 2015;44(D1):20–26. https://doi.org/10.1093/nar/gkv1352.
20. Alyass A, Turcotte M. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics.* 2015;8(1):1–12. https://doi.org/10.1186/s12920-015-0108-y.
21. Holzinger A. Knowledge discovery and interactive data mining in bioinformatics—state-of-the-art, future challenges, and research directions. *BMC Bioinf*. 2014;15(Suppl 6):S1–S9. https://doi.org/10.1186/1471-2105-15-S6-I1.
22. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016;17:1–19. https://doi.org/10.1186/s13059-016-0881-8.
23. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics*. 2015;16:1–12. https://doi.org/10.1186/1471-2164-16-S1-S2.
24. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015;16(6):321–332. https://doi.org/10.1038/nrg3920.
25. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet.* 2011;12(2):87–98. https://doi.org/10.1038/nrg2934.
26. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008;18(9):1509–1517. http://www.genome.org/cgi/doi/10.1101/gr.079558.108.
27. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470–476. https://doi.org/10.1038/nature07509.
28. Denoeud F, Aury JM, Da Silva C, et al. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* 2008;9:R175. https://doi.org/10.1186/gb-2008-9-12-r175.
29. Maher CA, Kumar-Sinha C, Cao X, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009;458(7234):97–101. https://doi.org/10.1038/nature07638.
30. Indira DNVSL, Kumar R, Gunavathi C, Reddy SK. Improved artificial neural network with state order dataset estimation for brain cancer cell diagnosis. *Biomed Res Int.* 2022;16(2022):1–12. https://doi.org/10.1155/2023/9842518.
31. Winchester LM, Harshfield EL, Williamson J, Nevado-Holgado AJ, Sims R, Escott-Price V. Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia. *Alzheimers Dement.* 2023;19:1–12. https://doi.org/10.1002/alz.13390.
32. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013;41(D1). https://doi.org/10.1093/nar/gks1193. D991-5.
33. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011;1(39):D19–D21. https://doi.org/10.1093/nar/gkq1019.
34. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources, and tools. Database. 2020 Aug 6; 2020: baaa062. Doi: 10.1093/database/baaa062.
35. Duan Y, Li X, Fan Y, Zhu Y, Dai Y, Zhou D. The correlation of ELP4-PAX6 with rolandic spike sources in idiopathic rolandic epilepsy syndromes. *Front Neurol.* 2021;9(12):1–9. https://doi.org/10.3389/fneur.2021.643964.
36. Pai B, Tom-Goh J, Eom K, Kyeong D. High-resolution transcriptomics informs glial pathology in human temporal lobe epilepsy. *Acta Neuropathol Commun.* 2022;10(1):1–19. https://doi.org/10.1186/s40478-022-01453-1.
37. Kjær C, Bager G, Olsen K. Transcriptome analysis in patients with temporal lobe epilepsy. *Brain*. 2019;142(10):1–14. https://doi.org/10.1093/brain/awz265.
38. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27(6):863–864. https://doi.org/10.1093/bioinformatics/btr026.
39. Sewe SO, Sanjai G, Lajunen HM, Hackl T, Harris RM. Trimming and validation of Illumina short reads using trimmomatic, trinity assembly, and assessment of RNA-Seq data. *Plant Bioinformatics.* 2022;16:211–232. https://doi.org/10.1007/978-1-0716-2067-0_11.
40. Wen G. A simple process of RNA-sequence analyses by Hisat2, Htseq, and DESeq2. *Proc Int Conf Biomed Eng Bioinformatics.* 2017;14:11–15. https://doi.org/10.1145/3143344.3143354.
41. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656.
42. Liu S, Wang Z, Su Y, Zhou Y, Zheng H. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *JoVE (J Vis Exp)*. 2021;175, e62299. https://doi.org/10.3791/62528.
43. Wang L, Xu Y, He W, Luo H. RNA-seq assistant: machine learning-based methods to identify more transcriptionally regulated genes. *BMC Genomics*. 2018;20(19):1–12. https://doi.org/10.1186/s12864-018-4932-2.
44. Ramzan M. Comparing and evaluating the performance of WEKA classifiers on critical diseases. 1st India Int Conf Inf Process (IICIP). 2016; 1-4. Doi: 10.1109/IICIP.2016.7975309.
45. Naik A, Sahu L. Correlation review of classification algorithms using data mining tools: WEKA, RapidMiner, Tanagra, Orange, and Knime. *Procedia Comput Sci.* 2016;85:662–668. https://doi.org/10.1016/j.procs.2016.05.251.
46. Reddy TR, Venkatesh B. Gender prediction in author profiling using ReliefF feature selection algorithm. Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA. 2018 Apr 11; 169-76. Doi: 10.1007/978-981-10-7566-7_18.
47. Kumar Y, Gopal G. Analysis of parametric & non-parametric classifiers for classification technique using WEKA. *Int J Inf Technol Comput Sci (IJITCS)*. 2012;4(7):43–49. https://doi.org/10.5815/ijitcs.2012.07.06.
48. Arora T, Dhiman R. Correlation-based feature selection and classification via regression of segmented chromosomes using geometric features. *Med Biol Eng Comput.* 2016;54(5):733–745. https://doi.org/10.1007/s11517-016-1553-2.
49. AL-Taie RRK, Jassim B. Analysis of WEKA data mining algorithms Bayes Net, Random Forest, MLP, and SMO for heart disease prediction: a case study in Iraq. *Int J Electr Comput Eng (IJECE)*. 2021;11(6):5229–5239. https://doi.org/10.11591/IJECE.V11I6.PP5229-5239.
50. Rajesh P, Manikandan A. A comparative study of data mining algorithms for decision tree approaches using WEKA tool. *Adv Nat Appl Sci.* 2017;11(7):230–243.
51. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. Proc 19th ACM SIGKDD Int Conf Knowl Discov Data Min. 2013 Aug 11; 847-55. doi: 10.1145/2487575.2487629.
52. Beunza JJ, Puertas E, García-Ovejero E, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *J Biomed Inform.* 2019;97, 103257. https://doi.org/10.1016/j.jbi.2019.103257.
53. Gupta A, Pratap R. Determining accuracy rate of artificial intelligence models using Python and R-Studio. *3rd Int Conf Adv Comput Commun Control Netw.*. 2021:889–894 https://doi.org/10.1109/ICAC3N53548.2021.9725687.
54. Patidar R, Mehta S. An integration of geospatial and machine learning techniques for mapping groundwater potential: a case study of the Shipra River Basin. *India. Arab J Geosci.* 2021;3(14):1–16. https://doi.org/10.1007/s12517-021-07871-0.
55. Sánchez-Baizán N, Rojas L, Moreno M. Improved biomarker discovery through a plot twist in transcriptomic data analysis. *BMC Biol.* 2022;24(20):1–26. https://doi.org/10.1186/s12915-022-01398-w.
56. Singal DP, Liao J. Genetic basis for rheumatoid arthritis. *Arch Immunol Ther Exp (warsz).* 1999;47(5):307–311.
57. Shastry BS. SNPs in disease gene mapping, medicinal drug development and evolution. *J Hum Genet.* 2007;52(10):871–880. https://doi.org/10.1007/s10038-007-0200-z.
58. Wang M, Wang J, Li X, et al. Genome-wide identification of RNA modification-related single nucleotide polymorphisms associated with rheumatoid arthritis. *BMC Genomics*. 2023;27(24):1–12. https://doi.org/10.1186/s12864-023-09227-2.

59. Wu R, Liu G, Yang L, Gao Y, Luo G. A novel m6A reader Prrc2a controls oligodendroglial specification and myelination. *Cell Res.* 2018;28(1):23–41. https://doi.org/10.1038/s41422-018-0113-8.
60. Karadag N, Ates A. Identification of novel genomic risk loci shared between common epilepsies and psychiatric disorders. *Brain.* 2023;146(2):1–14. https://doi.org/10.1093/brain/awad038.
61. Zhang Y, Chen K, Sloan SA, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci.* 2014;34(36):11929–11947. https://doi.org/10.1523/JNEUROSCI.1860-14.2014.
62. Zhang J, Chen M-J, Xu C-X, Fang F, Yu Y. Common genetic variants in PRRC2A are associated with both neuromyelitis optica spectrum disorder and multiple sclerosis in the Han Chinese population. *J Neurol.* 2021;268(2):506–515. https://doi.org/10.1007/s00415-020-10184-z.
63. Tang L, Wang X. The role of N6-methyladenosine modification in rodent models of neuropathic pain: from the mechanism to therapeutic potential. *Biomed Pharmacother.* 2023;166, 115179. https://doi.org/10.1016/j.biopha.2023.115398.
64. Liu X, Zhao Y, Li F, Sun X, Wang H. PRRC2A promotes hepatocellular carcinoma progression and associates with immune infiltration. *J Hepatocell Carcinoma.* 2021;1(8):1495–1511. https://doi.org/10.2147/JHC.S337111.
65. Su H, Wang Y, Liu Y, Zhang Z, Chen L. RNA m6A methylation regulators multi-omics analysis in prostate cancer. *Front Genet.* 2021;12:1–15. https://doi.org/10.3389/fgene.2021.768041.
66. Yin Y, Feng W, Zhang X, Huang Y, Li L. Immunosuppressive tumor microenvironment in the progression, metastasis, and therapy of hepatocellular carcinoma: from bench to bedside. *Exp Hematol Oncol.* 2024;13:1–20. https://doi.org/10.1186/s40164-024-00539-x.
67. Liu N, Wang X, Zhang Y, Li Z. MicroRNA-206 promotes the recruitment of CD8+ T cells by driving M1 polarization of Kupffer cells. *Gut.* 2022;76(10):1642–1655. https://doi.org/10.1136/gutjnl-2021-324170.
68. Sekine T, Perez-Potti A, Nguyen S, Gorin J-B, Wu VH, Gostick E, et al. TOX is expressed by exhausted and polyfunctional human effector memory CD8+ T cells. Sci Immunol. 2020 Jul 3; 5(49): eaba7918. Doi: 10.1126/sciimmunol.aba7918.
69. Hudson WH, Gensheimer J, Xu J, et al. Proliferating transitory T cells with an effector-like transcriptional signature emerge from PD-1+ stem-like CD8+ T cells during chronic infection. *Immunity*. 2019;51(6):1043–1058. https://doi.org/10.1016/j.immuni.2019.11.002.
70. Chen Y, Sun M, Zhang H, Li Y, Liu X. Genome-wide identification of N6-methyladenosine-associated SNPs as potential functional variants for type 1 diabetes. *Front Endocrinol (Lausanne).* 2022;16(13):1–12. https://doi.org/10.3389/fendo.2022.913345.
71. Toker A, Dufner M. AKT/PKB signaling: navigating the network. *Cell.* 2017;169(3):381–405. https://doi.org/10.1016/j.cell.2017.04.001.
72. Pavlov VA, Chavan SS, Tracey KJ. Molecular and functional neuroscience in immunity. *Annu Rev Immunol.* 2018;26(36):783–812. https://doi.org/10.1146/annurev-immunol-042617-053158.