ISSN 1870-4069

Machine Learning for Biomarker Identification in Ischemic Stroke Patients

Rodolfo Betanzos Cerqueda[1,3], Noé Macías Segura[2], Dulce Martinez-Peon[1,3],
Rodrigo Sánchez Zavala[1], Fernando Góngora-Rivera[2],
Christian Quintus Scheckhuber[4]

[1] Tecnológico Nacional de México,
División de Estudios de Posgrado e Investigación,
Mexico

[2] Universidad Autónoma de Nuevo León, Monterrey,
Facultad de Medicina,
Mexico

[3] Tecnológico Nacional de México,
Departamento de Ingeniería Eléctrica y Electrónica,
Mexico

[4] Escuela de Ingeniería y Ciencias,
Tecnológico de Monterrey,
Mexico

{dd244881482, dulce.mp} @ nuevoleon.tecnm.mx

Abstract. Stroke is a medical condition that increasingly affects younger people around the world. Biomarkers are a helpful tool for diagnosing medical conditions based on genetic data. Typically, bioinformatics is used to identify which genes are candidates for a biomarker; however, this tool depends on linearly based algorithms. Machine Learning algorithms have been demonstrated to be useful for the detection of clue genes for certain diseases and medical conditions. In this work, we used a database from GEO containing data on stroke patients. We apply three algorithms—Support Vector Machines, Extreme Gradient Boosting, and Random Forest—to identify genes whose expression has a meaningful difference between the control group and stroke patients. The obtained results reveal sixteen genes: SVIL, C5AR1, MAX, KIF1B, ACOX1, PLXDC2, TNFRSF17, DOCK8, PHTF1, TRIB1, CREBBP, NPEPPS, RGS2, FAM108A3, ST8SIA4, and CD163.

Keywords: Gene expression, molecular pathways, biomarkers.

1 Introduction

Stroke is one of the leading causes of incapacity and death around the world. Each year, 12 million patients present this medical condition, and 6.5 million patients dies [1]. In Mexico, there are 118 stroke patients per 100,000 inhabitants, and 170,000 cases are reported annually, with around 36,000 deaths [2]. Among the potential risks for stroke

pp. 109–121; rec. 2025-03-16; acc. 2025-05-07

are diabetes, hypertension, obesity, and smoking [3]. Early detection is crucial to improving the prognosis. In this sense, recent tools like biomarkers are used to complement existing tools such as neuroimaging [4].

Gene expression through microarray systems has characterized neurological diseases and immune disorders [5]. It contains information about the RNA that is obtained from blood. Bioinformatic tools typically process this information. However, the data type provided by the microarrays has been tackled recently, with Machine Learning algorithms showing prominent results [6].

In this work, we used a public data set of Ischemic stroke patients and applied three ML algorithms, Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and Random Forest (RF), to identify genes with the highest expression in patients against a control group.

Recent advances in transcriptomics and machine learning have demonstrated the ability of ML algorithms to uncover complex biomarker signatures in ischemic stroke. For example, O'Connell et al. (2017) used ML to identify a peripheral blood gene-expression signature that diagnoses ischemic stroke with over 90 % sensitivity and specificity. More recently, Liu et al. (2024) applied Random Forest, Support Vector Machine (SVM), and XGBoost to coagulation-related gene expression, highlighting ACTN1, F5, and JMJD1C as robust diagnostic markers. These studies illustrate that ensemble methods and gradient-boosting frameworks excel at modeling nonlinear gene interactions in high-dimensional data. In a complementary proteomic approach, Dargazanli et al. (2020) leveraged SVM to differentiate cardioembolic from atherothrombotic thrombi, reinforcing the versatility of ML in stroke biomarker discovery. Unsupervised techniques—such as autoencoders and clustering—have also revealed hidden molecular subtypes among stroke patients (Liu et al., 2022; Burrello et al., 2022), paving the way for precision-medicine strategies. Collectively, this body of work justifies our selection of Random Forest, XGBoost, and SVM: these algorithms bring complementary strengths in handling noise, correcting misclassifications iteratively, and maximizing class separation, respectively, and each has demonstrated proven success in prior stroke-related omics investigations.

2 Materials and Methods

A dataset obtained from the Gene Expression Omnibus (GEO) database was utilized to reanalize. Methods employed are described and detailed in this section including the ML algorithms, to identify biomarkers and examine differential gene expression in patients diagnosed with ischemic stroke disease, and the statistical modelling approaches used to evaluate, see Fig. 1.

2.1 Data Acquisition and Preprocessing

We used the GSE16561 series (Accession: GSE16561) from NCBI's Gene Expression Omnibus, which comprises 63 peripheral whole-blood RNA samples—39 acute ischemic stroke patients (MRI-confirmed, >18 years, collected in PAXgene Blood RNA tubes) and 24 neurologically healthy controls—profiled on the Illumina HumanRef-8 v3.0 expression beadchip (GPL6883). The dataset comprises 39 patients

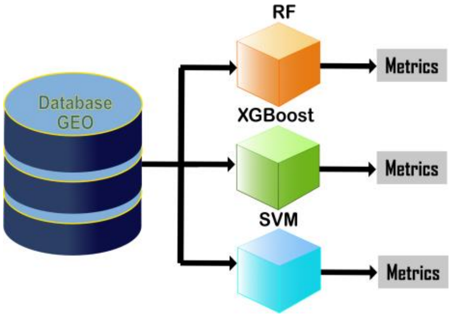

Fig. 1. Block diagram of the processing for gene data.

with ischemic stroke (IS) and 24 control individuals [5,7,8]. The process involved the removal of missing values and the transformation of the data into a numerical structure.

The analysis did not involve filtering out low-variance genes; therefore, all genes were included. This approach is justified by the aim of the study, which was to directly assess the ability of supervised models to identify the most relevant features. Applying an initial variance-based filtering step could potentially remove genes that, despite exhibiting low overall variability, may still be discriminative for classification purposes.

In this study, the dataset was already normalized in its original format; therefore, no additional normalization procedures were implemented during preprocessing. The assessment of gene expression is constrained by the high dimensionality of the data, as a single experiment may encompass tens of thousands of genes as predictive variables. To address this challenge, feature selection methods based on supervised machine learning models were employed, enabling the reduction of the gene set to those contributing the most relevant information for the classification of patients with stroke.

2.2 Classification Models

Three representative classification approaches were implemented.

Random Forest (RF): This ensemble model utilizes multiple decision trees to generate predictions. Gene importance is assessed by measuring the contribution of each gene to the overall improvement in classification accuracy.

XGBoost (Extreme Gradient Boosting): This advanced tree-based model builds predictive estimators through iterative boosting. Gene importance is evaluated based on how significantly each gene contributes to the quality of the model's decision-making during classification.

Support Vector Machine (SVM) with Linear Kernel: This algorithm identifies the optimal hyperplane that separates patients with cerebral infarction from control subjects. The relevance of each gene is determined by analyzing the magnitude of its contribution to this separation.

These models were selected due to their distinct strategies for capturing complex relationships within the data. RF is effective when genes interact in intricate, non-linear ways that do not conform to simple patterns. It is a robust model and performs well in noise or data imperfections. XGBoost is an advanced model that incrementally improves accuracy by correcting errors at each stage of learning. This iterative

Table 1. Comparative strengths and gene importance criteria of the classification models used in this study.

Model	Main Strengths	Gene importance criteria
Random Forest	-Handles nonlinear interactions between genes. -Robust to noise and overfitting -Works well with high-dimensional data	Mean decrease in impurity (e.g., Gini): measures how much each gene reduces classification error across trees.
XGBoost	-Highly accurate through iterative boosting -Efficient and scalable -Corrects misclassifications at each stage of training	Gain: evaluates the improvement in classification accuracy when a gene is used to split decision tree nodes
Linear SVM	-Finds optimal linear separation between classes -Performs well with high-dimensional, sparse data -Easy to interpret in linear form	Magnitude of model coefficients: genes with larger weights contribute more to the separation hyperplane

enhancement makes it highly efficient and capable of achieving high classification accuracy. Linear SVM identifies the optimal combination of genes that clearly distinguishes between patients with cerebral infarction and healthy individuals.

Specifically, Random Forest was selected because it handles high-dimensional data well and can model complex, nonlinear interactions among genes without overfitting, making it robust to noisy microarray measurements. XGBoost was included for its state-of-the-art gradient-boosting framework, which iteratively corrects classification errors and incorporates regularization to prevent overfitting, delivering both high accuracy and scalability on large feature sets. Finally, a linear Support Vector Machine was employed due to its proven effectiveness in high-dimensional, sparse settings like gene expression, where it finds the optimal separating hyperplane and yields easily interpretable feature weights. Together, these methods represent complementary approaches—ensemble averaging, boosting, and maximum-margin classification—that ensure a comprehensive evaluation of biomarker relevance across different modeling paradigms.

Table 1 provides a summary of the key strengths of each classification model used in this study.

2.3 Feature Importance Estimation

Each model evaluates gene importance according to its underlying criteria. RF determines the relevance of each gene by assessing how much it contributes to improving classification accuracy at each stage. If the removal of a gene leads to a significant drop in accuracy, the gene is considered essential. XGBoost measures gene

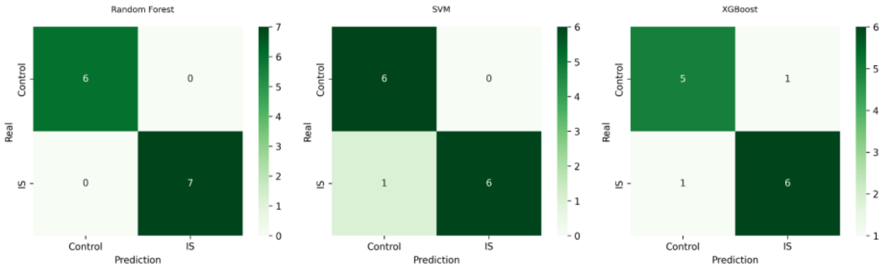

Fig. 2. Confusion matrix resume.

importance by analyzing how frequently a gene is used at critical decision points in the model. A gene repeatedly selected and enhancing prediction performance is deemed highly informative. Linear SVM identifies the optimal boundary for separating patients from controls and assigns a weight to each gene. Genes with higher absolute weights exert a more significant influence on the classification outcome.

To ensure the validity of the findings, potential biomarkers were defined as those genes that appeared among the top 100 ranked features in at least two or more models. This criterion suggests that their importance is not tied to a single method but demonstrates consistency across multiple analytical approaches.

2.4 Metrics

To evaluate the classification performance of each machine learning model, confusion matrices were generated for the top 100 most important genes selected by each method. These matrices illustrate the number of true positives, true negatives, false positives, and false negatives in distinguishing ischemic stroke (IS) patients from control group.

3 Results

Fig. 2 to Fig. 6 show the results obtained.

The RF model achieved perfect classification performance, correctly identifying all control and IS samples (6 true negatives, 7 true positives), resulting in 100% accuracy. The SVM with a linear kernel also showed high performance, correctly classifying all control samples and misclassifying only one IS case as control, achieving an overall accuracy of 92.9%. The XGBoost model correctly identified 5 out of 6 control subjects and 6 out of 7 IS patients, with one false positive and one false negative, yielding an overall accuracy of 85.7%.

These results demonstrate the high discriminative power of the selected gene subsets across different classifiers, with Random Forest showing the most robust performance in this experimental setup.

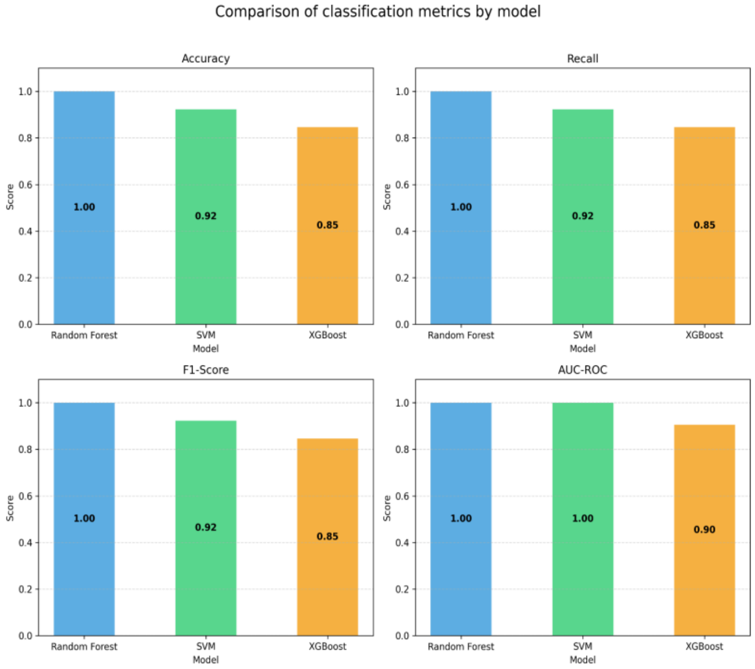

Fig. 3. Comparison of classification metrics by model.

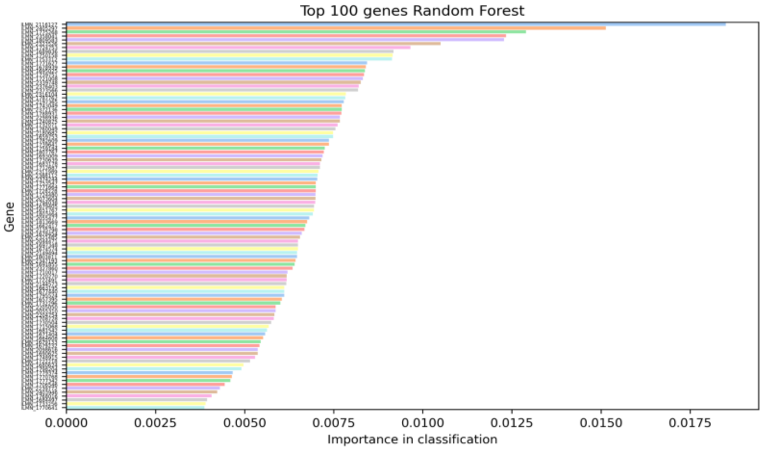

Fig. 4. Top 100 most important genes identified by the Random Forest model.

To complement the confusion matrix analysis, additional classification metrics were calculated for each model, including precision, recall, F1-score, and area under the ROC curve (AUC-ROC). As shown in the table below, Random Forest achieved perfect scores across all metrics, while SVM and XGBoost also demonstrated strong performance with slightly lower values in class-specific precision and recall. These

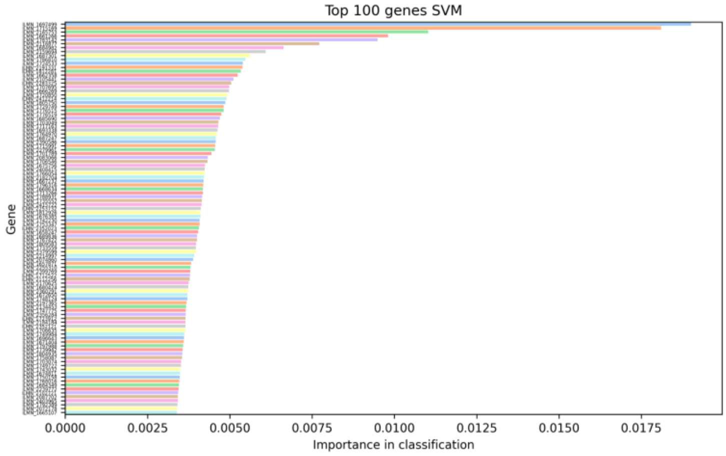

Fig. 5. Top 100 most important genes identified by the SVM model.

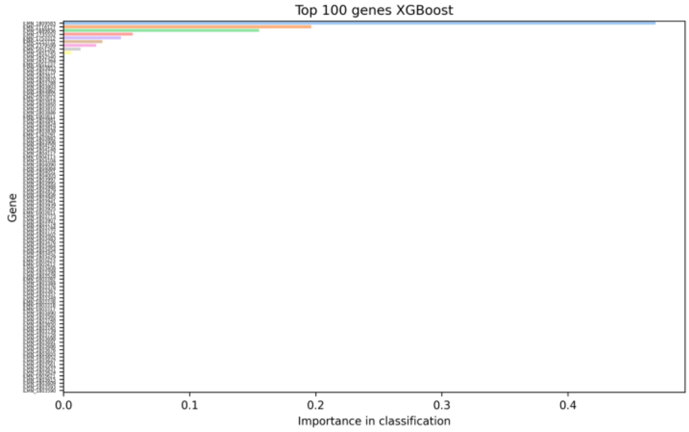

Fig. 6. Top 100 most important genes identified by the XGBoost model.

metrics reinforce the conclusions drawn from the confusion matrices, confirming that Random Forest provided the most robust classification among the three methods evaluated.

3.1. Gene Importance Comparison

The relative importance of genes in classification was analyzed across the three machine learning models: Random
Forest, Support Vector Machine (SVM), and XGBoost. Each model ranked the top 100 genes according to its internal criteria for feature relevance. The Random Forest model distributed importance more evenly across many genes, suggesting a broader contribution of features to classification decisions.

In contrast, the SVM model exhibited a steep decline in importance values, with only a small subset of genes carrying significantly higher weights, indicating a more selective dependence on a limited number of features.

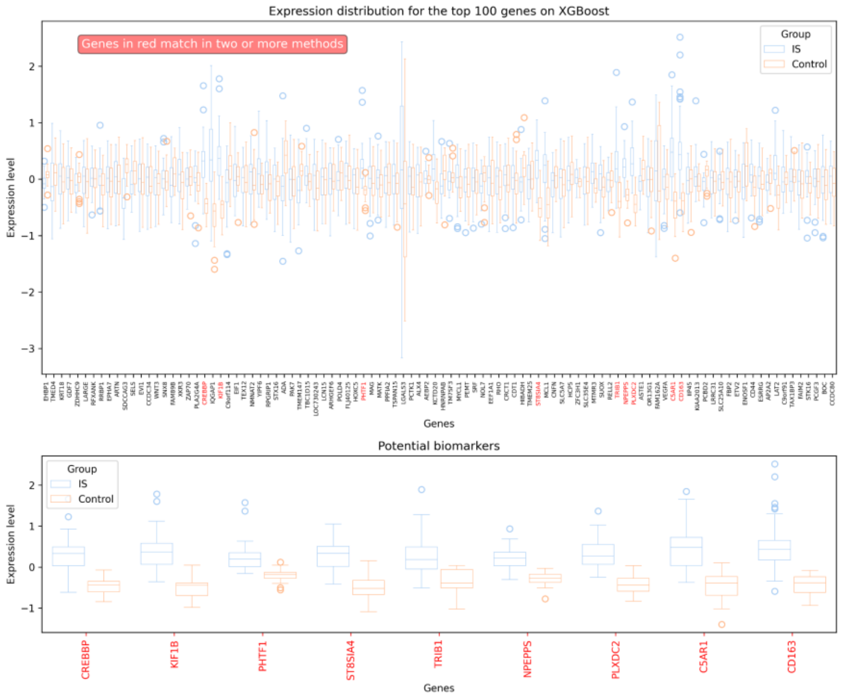

Fig. 7. Gene expression distribution for top 100 genes identified by XGBoost. Genes highlighted in red are shared by two or more models and considered potential biomarkers.

XGBoost demonstrated a highly concentrated importance profile, where only a few genes dominated the classification decision-making process. The top-ranked gene in the XGBoost model contributed disproportionately more to the overall model accuracy compared to the rest.

These patterns reflect the intrinsic characteristics of each algorithm. Random Forest benefits from ensemble averaging and tends to distribute importance broadly. SVM, being a linear classifier, identifies the most discriminative directions in the feature space.

XGBoost, as a boosting-based method, favors feature that provide the highest gain at each step of model construction. A comparison of gene rankings among models also revealed overlapping genes, which reinforces the robustness of the identified biomarkers and supports their biological relevance in the context of ischemic stroke classification.

3.2 Selection of Genes Repeated on at least Two Methods

To enhance the reliability of biomarker discovery, genes that appeared among the top 100 features in at least two out of the three models (Random Forest, SVM, and XGBoost) were considered potential biomarkers. This criterion ensures that gene

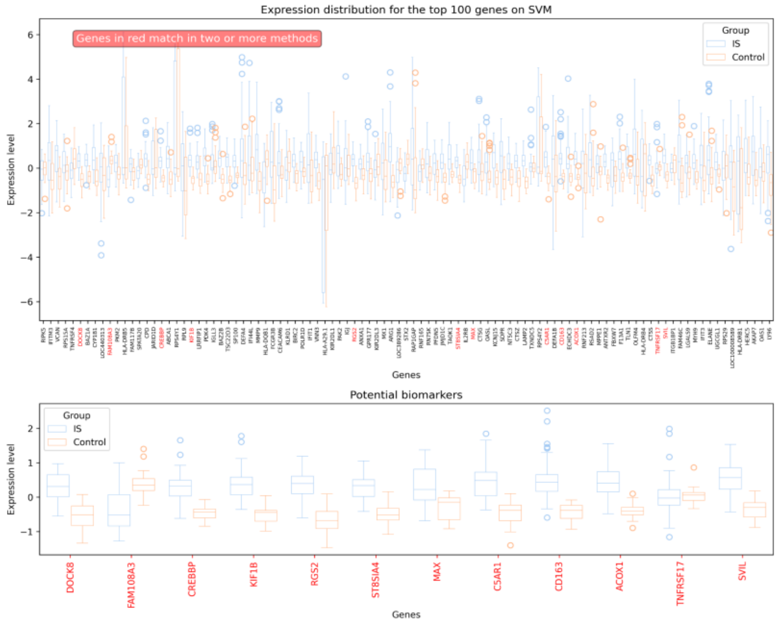

Fig. 8. Gene expression distribution for top 100 genes identified by SVM. Potential biomarkers shared across models are labeled in red.

importance is consistent across multiple learning paradigms, minimizing model-specific biases.

3.2.1 XGBoost

In the XGBoost model, several genes marked in red were identified as recurring across multiple methods. These genes showed consistent expression differences between ischemic stroke (IS) patients and control subjects. The bottom plot highlights the expression distributions for those overlapping genes, revealing clear group separation in many cases.

3.2.2 Support Vector Machine (SVM)

The SVM model revealed a broader distribution of gene expression values, and many of the recurring genes also showed distinctive expression profiles. Notably, genes such as CD163, CREBBP, and C5AR1 demonstrated clear upregulation or downregulation patterns in the IS group.

3.2.3. Random Forest

Random Forest provided a more balanced view, with overlapping genes such as PLXDC2, RGS2, TRIB1, and SVIL showing distinguishable expression levels between

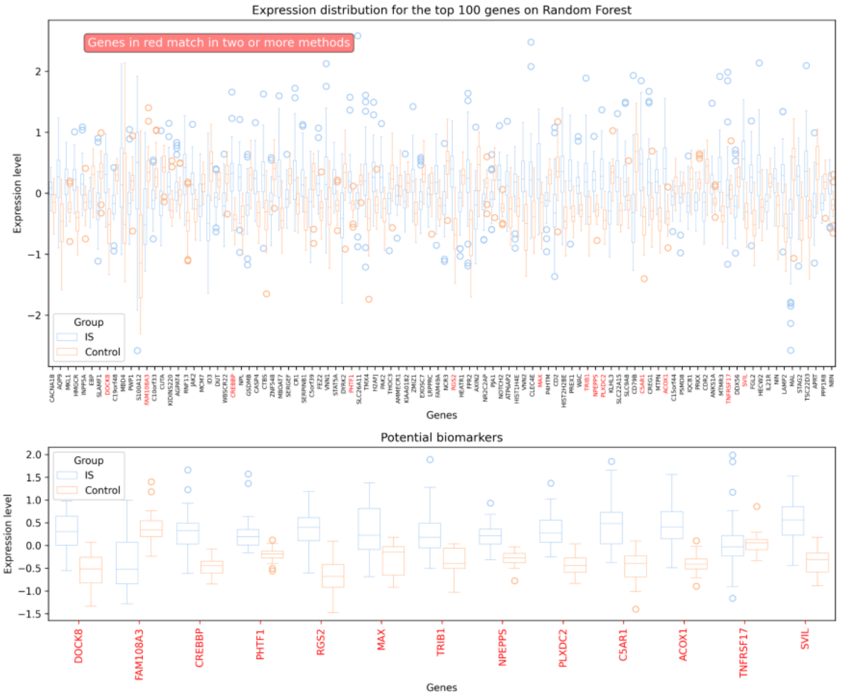

Fig. 9. Gene expression distribution for top 100 genes identified by Random Forest. Highlighted genes in red appeared in at least two models.

IS and control groups. These expression profiles further support their candidacy as robust biomarkers.

Genes that appeared in two or more methods were identified and considered potential biomarkers in this study.

The selected genes are: SVIL, C5AR1, MAX, KIF1B, ACOX1, PLXDC2, TNFRSF17, DOCK8, PHTF1, TRIB1, CREBBP and NPEPPS.

3.3. Molecular Pathway Analysis

To gain deeper insight into the biological functions and interactions of the identified genes, a network-based molecular pathway analysis was performed using the GeneMANIA platform. This tool integrates data from multiple sources to predict gene-gene interactions based on co-expression, co-localization, physical interactions, and genetic interactions.

The resulting network, shown in Fig. 10, reveals a densely interconnected structure among the selected genes. Notably, CD163, TRIB1, CREBBP, and C5AR1 emerge as central nodes, suggesting that they play pivotal regulatory roles in ischemic stroke pathology. CD163, a scavenger receptor expressed in monocytes and macrophages, contributes to anti-inflammatory responses following tissue injury. TRIB1 participates in lipid metabolism and macrophage polarization, processes closely linked to vascular

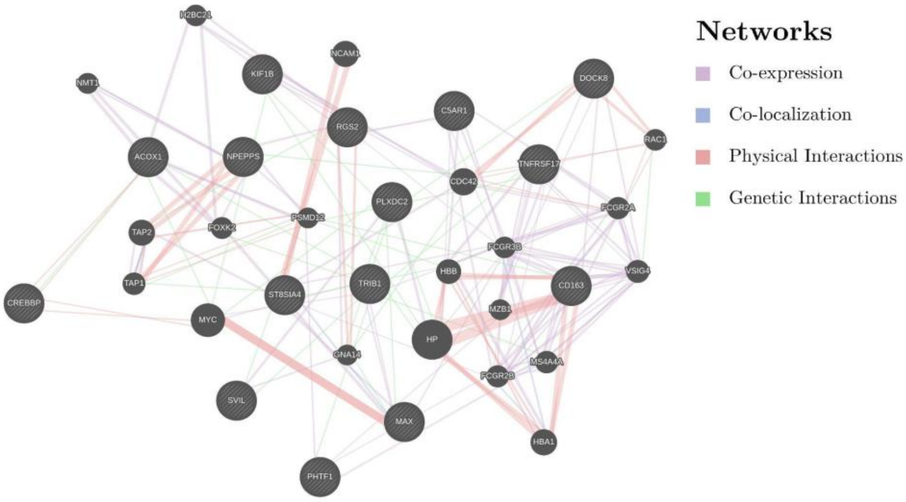

Fig. 10. Gene interaction network generated using GeneMANIA, illustrating functional associations among the selected candidate biomarkers for ischemic stroke. Edge colors represent different types of interactions: co-expression (purple), co-localization (blue), physical interactions (red), and genetic interactions (green).

inflammation. CREBBP, known as a transcriptional coactivator, modulates the expression of genes involved in cell survival and immune regulation. Meanwhile, C5AR1 acts as a receptor for complement component C5a and is strongly implicated in neuroinflammation and ischemia-reperfusion injury.

Edge colors in the network indicate different types of interactions: co-expression (purple edges) reflects genes expressed simultaneously under similar conditions, suggesting shared regulatory mechanisms; co-localization (blue edges) indicates that genes are located within the same cellular compartments, implying potential cooperation in localized biological processes; physical interactions (red edges) represent direct binding between protein products; and genetic interactions (green edges) highlight functional interdependencies inferred from genetic perturbation studies.

4 Discussion and Conclusions

This study enabled the identification of potential biomarkers with diagnostic and prognostic relevance for ischemic stroke. Machine learning models facilitated the evaluation of gene expression data, and their integration with molecular pathway analysis provided a more comprehensive perspective on the underlying biological mechanisms.

The findings suggest that the implementation of machine learning methodologies not only enhances the accuracy of biomarker detection but also simplifies the biological interpretation of genes involved in the pathology. Future research will aim to expand the analysis by incorporating a larger patient cohort and exploring model interpretation techniques to optimize the understanding of the relationship between the identified genes and the disease.

In addition, the integration of other omics data types—such as proteomics or metabolomics—is envisioned to achieve a more holistic understanding of ischemic stroke biology and to support the development of more accurate diagnostic tools.

The presence of multiple interaction types among these key genes reinforces their biological relevance and underscores a cooperative molecular framework underlying ischemic stroke. This network-based systems biology approach points to CD163, TRIB1, CREBBP, and C5AR1 as promising biomarkers and potential targets for therapeutic intervention.

References

1. GBD Stroke Collaborators: Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurology, 20(10), pp. 795–820 (2021) doi: 10.1016/S1474-4422(21)00252-0.
2. Jiménez Muñiz, V.E.: Accidente cerebrovascular, cuarta causa de muerte en México en mayores de 45 años. Universidad de Guadalajara. Accidente cerebrovascular, cuarta causa de muerte en México en mayores de 45 años, Universidad de Guadalajara (2024).
3. Mitchell, A.B., Cole, J.W., McArdle, P.F., Cheng, Y.C., Ryan, K.A., Sparks, M.J., Kittner, S.J.: Obesity increases risk of ischemic stroke in young adults. Stroke, 46(6), pp. 1690–1692 (2015)
4. Mendioroz Iriarte, M., Cuadrado Godia, E., & Montaner Villalonga, J.: Biomarcadores plasmáticos en la enfermedad vascular cerebral isquémica [Plasma biomarkers in ischemic cerebral vascular disease]. Hipertensión, 26(6), pp. 266–274 (2009) doi: 10.1016/j.hipert.2008.07.001.
5. Barr, T.L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A., Matarin, M.: Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. Neurology, 75(11), pp. 1009–1014 (2010)
6. Tabl, A.A., Alkhateeb, A., ElMaraghy, W., Rueda, L., Ngom, A.: A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Frontiers in genetics, 10, pp. 256 (2019)
7. O'Connell G.C, Treadway M.B, Petrone A.B, Tennant C.S, et al.: Peripheral blood AKAP7 expression as an early marker for lymphocyte-mediated post-stroke blood brain barrier disruption. Sci Rep 7(1), pp. 1172, PMID: 28446746 (2017)
8. O'Connell G.C, Petrone A.B, Treadway M.B, Tennant C.S et al.: Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. NPJ Genom Med 2016(1), pp. 16038, PMID: 29263821 (2016)
9. O’Connell, G.C., Chantler, P.D., Barr, T.L.: Stroke-associated pattern of gene expression previously identified by machine-learning is diagnostically robust in an independent patient population. Genomics Data, 14, pp. 47–52 (2017) doi: 10.1016/j.gdata.2017.08.006.
10. Liu, J., Si, Z., Liu, J., Zhang, X., Xie, C., Zhao, W., Wang, A., Xia, Z.: Machine learning identifies novel coagulation genes as diagnostic and immunological biomarkers in ischemic stroke. (2024) doi: 10.18632/aging.205706.
11. Dargazanli, C., Zub, E., Deverdun, J., Decourcelle, M., De Bock, F., Labreuche, J., Lefèvre, P.H., Gascou, G., Derraz, I., Riquelme Bareiro, C., Cagnazzo, F., Bonafé, A., Marin, P., Costalat, V., Marchi, N.: Machine Learning Analysis of the Cerebrovascular Thrombi Proteome in Human Ischemic Stroke: An Exploratory Study. Frontiers in Neurology, 11, pp. 575376 (2020) doi: 10.3389/fneur.2020.575376.

12. Liu, J., Chou, E.L., Lau, K.K., Woo, P.Y.M., Li, J., Chan, K.H.K.: Machine Learning Algorithms Identify Demographics, Dietary Features, and Blood Biomarkers Associated with Stroke Records. Journal of the Neurological Sciences, 440, pp. 120335 (2022) doi: 1016/j.jns.2022.120335.
13. Burrello, J., Burrello, A., Vacchi, E., Bianco, G., Caporali, E., Amongero, M., Airale, L., Bolis, S., Vassalli, G., Cereda, C. W., Mulatero, P., Bussolati, B., Camici, G. G., Melli, G., Monticone, S., Barile, L.: Supervised and unsupervised learning to define the cardiovascular risk of patients according to an extracellular vesicle molecular signature. Translational Research, 244, pp. 114-125 (2022) doi: 10.1016/j.trsl.2022.02.005.