Received: 31 October 2024 | Revised: 5 June 2025 | Accepted: 18 July 2025 | Published online: 4 August 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–13
DOI: 10.47852/bonviewMEDIN52024698

BON VIEW PUBLISHING

Gene Signatures for Autism Classification: Mining Biological Markers for Autism from Gene Expression Data

Shomona Gracia Jacob[1,*]

[1]*School of ICT, Bahrain Polytechnic, Kingdom of Bahrain*

Abstract: Autism spectrum disorders are reported to be one of the most intriguing neurodegenerative conditions, while disinterring the possible causes of this malady has been a topic of intense research in the recent past. Many studies trace the origin of autism to gene mutations. However, it has been reported that analyzing hundreds of genes present in the human body led to extensive use of resources in terms of expertise, capital, and time. This in turn paved the way for computational investigations on gene expression data, which also proved to be a challenging task owing to the momentous number of attributes and the relatively low number of instances that were available to train the machine learning models. This research work thus explores the use of automated machine learning, deep learning, and traditional machine learning models to detect possible gene signatures that play the most contributory role in characterizing the presence of autism. The results suggest that the Bayesian classifier model fused with correlation feature filtering yielded higher accuracy, this being reported for the first time on this gene expression data. The proposed Bayesian machine learning model generated an accuracy of ∼87% with a minimal yet optimal gene signature that ranks a subset of 22 genes as significant gene markers from a total of 9454 genes.

Keywords: machine learning, Naive Bayes classification, feature selection, autism detection, gene markers

1. Introduction

Autism spectrum disorder (ASD) is identified as a neurocognitive degenerative condition that exhibits symptoms such as indifferent social interactions, weakened communication, and cyclic mode of actions [1]. It has also been reported in previous studies that genetic similarity among identical twins who belong to the same ecological environment and inherit similar parental chromosomes is high [2, 3].

Recent research on autism and the application of computational methods for the diagnosis of this condition has revealed the need to unearth its causes, such as the role of inheritance in autism, the most contributing gene, the most contributing mutation, environmental factors, exposure to abuse, and so on [4]. Most of the studies in the recent past have focused on real-time data collected from individuals restricted to a certain region. Some published reports have reported that ethnicity or familial history is not a major contributing factor to autism. Yet another challenging factor in autism research is identifying the kind of therapeutic measure to treat autistic children based on their age, cause of autism, socio-economic status, and educational standards of the family members [5].

Presently, gene biomarkers have not yet been precisely identified since manual processing and statistical analysis of the available patient data have proved to be a herculean task, thus inhibiting major breakthroughs in identifying ASD triggers from genetic information [6, 7]. Most studies have concentrated on the role of hereditary factors but have fallen short of the expected results owing to the complex structure of ASD genetics. Some reports claim that a collection of genes played a major role in triggering ASD, thereby augmenting the complexity of genome-based investigations [8–11].

This research work emphasizes the role of machine learning techniques in generating gene signatures that could lead us to the most contributing genes that can classify autism based on the gene expression values. Gene signature refers to the set of genes whose values can computationally detect the prevalence of autism with high accuracy. To portray the need for this study with suitable evidence, the recent research on the application of machine learning and data mining in the classification of autism is concisely tabulated here.

Based on the review of recent and related work in the classification of gene signatures for autism, it can be summarized based on the review in Table 1, to state that:

1) There is a lack of robust machine learning models and standardized databases that can predict the role of contributing factors for autism.
2) A comprehensive database of possible contributing factors from a social/medical perspective is unavailable.

*Corresponding author: Shomona Gracia Jacob, School of ICT, Bahrain Polytechnic, Kingdom of Bahrain. Email: shomona.jacob@polytechnic.bh

Table 1. Summary of recent and related work on genetic factors affecting autism

S.No	Reference	Objective	Findings	Inference
1.	Hameed et al. [11]	Identify discriminatory genes for autism.	CAPS2 was the most discriminative gene identified by the GBPSO-SVM algorithm with 86.3% accuracy but was a highly time-consuming process.	• No computational studies were done to substantiate the findings. • High level of preprocessing done on the data. • More precise evaluation metrics to be used for unbalanced datasets.
2.	Gunning et al. [12]	Review on the application of machine learning models to detect autism.	There was ample scope for improvement in the accuracy of the models.	• More work on generating and validating statistical models for predicting the association of rare gene variants and disease is needed.
4.	Brueggeman et al. [9]	Proposed an ensemble method, forecASD, which integrated brain gene expression, diverse brain network data, and existing gene-level predictors of autism association into an ensemble classifier. The score yielded by the classifier was treated as the evidence of each gene's involvement in the etiology of autism.	Only uses the TADA score to substantiate the results	• Does not record the evaluation metrics of the classifiers. • Works only on balanced data.
3.	Lin et al. [13]	Proposed an ML technique on the SFARI gene database to detect potential genetic causes for autism.	Gradient boosted tree model was recorded to yield the highest accuracy in detecting high-risk genes for autism.	• The method was tested only on the de novo mutation genes, so it does not explain inherited autism.
5.	Zaman et al. [14]	Detect genes that play a key role in cognitive function.	Genetic mutations can affect brain development in a fetus, leading to ASD	• Lack of methods to investigate the onset of ASD during gestation.
6.	Wu et al. [15]	Review on the significance of gene function in ASD onset.	Identify significant genes that trigger ASD in children.	• Lack of in vitro/in silico methods for early diagnosis of ASD.
7.	Rastagari et al. [16]	Proposed the FA_gene algorithm to find a minimal set of genes involved in autism.	Focused on multinomial classification that distinguished between different autism classes, namely, Asperger's syndrome, autism spectrum disorder, pervasive developmental disorder (PDD)-NOS (not otherwise specified), and control cases.	• No evaluation metrics were recorded to justify the findings with statistical significance.
8.	Gogate et al. [1]	The focus was on ASD genetics to detect ample variants across diverse ancestral backgrounds.	Computational investigations were not included in the study, and the work involved a molecular study and diagnosis.	• Unearths the need for computational methods to detect potential genetic markers for autism.
9.	Singh et al. [17]	A graph convolution network with logistic regression was used to identify potential genes based on the data retrieved from the PPI network.	The findings were compared with the SFARI database and the EAGLE framework to substantiate the derived results.	• More computational investigations using multimodal data and graph theory were suggested to identify the complex genetic interactions in autism.

3) Unbalanced dataset predictions have not utilized appropriate evaluation metrics to substantiate the findings with the required statistical significance.

The focus of this article is to identify the most discriminative gene signature for autism classification from the NCBI Geo Gene Expression Data [18] using machine learning approaches. The proposed model in this work is trained on the gene expression data comprising ~9500 genes and 146 records. The highly predictive gene signature comprises ~22 genes that generated a classification accuracy of ~87%, the highest reported thus far in the literature. The objective of this study is:

1) To identify the optimal set of genes that could direct further biological and genetic assays on the highly discriminative genes that could serve as biomarkers for autism.
2) Investigate the comparative performance of automated machine learning (AutoML) models, conventional machine learning, and deep learning models.
3) This is only the second reported study on this dataset, which is continuous, highly variant, with an extensive domain of values, and unbalanced. Hence, the authors propose to explore the optimization of hyperparameters that would yield the most accurate results in the classification of autism with the potential gene signatures.

The remaining sections of the manuscript are organized as follows: Materials and Methods, followed by Experimental Results, Discussions, and Conclusion.

2. Materials and Methods

The proposed methodology in this work is portrayed in Figure 1.

The gene expression data utilized to carry out this research were downloaded from the GEO expression database available at NCBI [18]. The data comprised 146 instances and 54,613 genes/attributes, the description of which is given in Figure 2. This dataset could be used only for binary classification as the target class was divided into the control class, containing 66 observations, and the autism class, comprising 80 records. This work highlights the results of the earlier work on the same dataset as reported by Hameed et al. [11], who had recorded a different number of samples in the respective classes, although they used the same source for data collection. Further details on the mode of data collection and the microarray experiments done have been described in Hameed et al. [11]. This work attempts to unveil the performance of traditional, neural network-based, and automated machine learning models in autism classification from gene expression data. This research work was focused on discovering whether a single gene or certain gene combinations (gene signatures) acted as potential ASD triggers.

The GEO expression dataset comprised ~9454 genes that were obtained after removing covariate features that had little bearing on the dependent variable. This dataset also contained information on the age of the child and the paternal and maternal ages of the parents. The author primarily investigated the role of parental age in genetic association and autism cause. However, there was ample missing data, and hence, the performance of machine learning algorithms in classification seemed out of line. Hence, the author placed emphasis only on the complete genetic data of the individual cases and evaluated the classifier's performance. It is evident from the statistical description of the dataset that there is an attribute-instance imbalance in the data, with the number of independent variables spanning ~ 9500 while the number of instances is restricted to 146.

This moved the author to assess the classifier performance based on the statistical metrics for unbalanced data, namely Mathew's correlation coefficient and balanced accuracy, both measures being reported for the first time in the literature. The 117_at gene has a domain spanning from a few hundred to many thousands, whereas the 244071_at gene has a domain restricted to 2-digit values. Such diverse domain distribution of continuous values makes it an arduous task for classification. This is one of

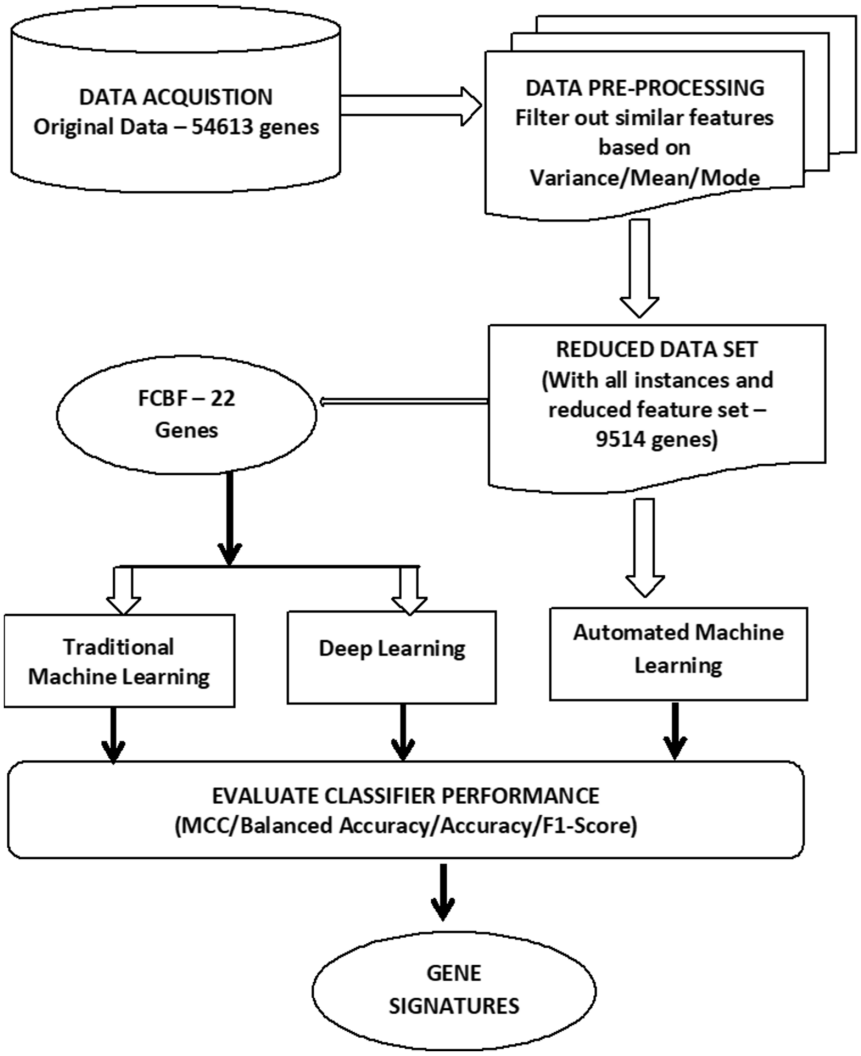


Figure 1. Proposed methodology for gene signature detection through machine learning

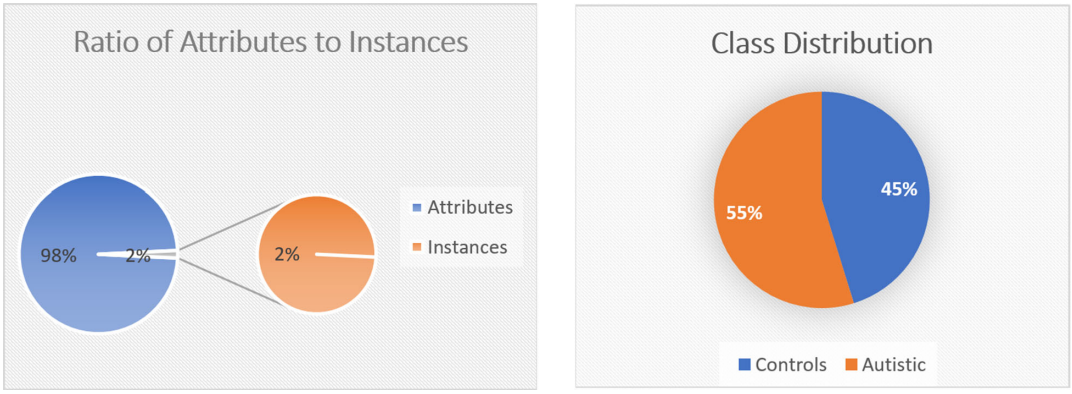


Figure 2. Data description of autism gene expression data – GEO database

Metrics	87100_at Numerical	91682_at Numerical	91703_at Numerical	91816_f_at Numerical	Class Categorical
min	9.983	33.003	304.018	9.575	
max	62.693	108.891	1892.690	46.465	
nanmean	21.378	46.898	814.957	20.088	
nanstd	8.068	8.531	263.044	5.452	
median	18.920	45.297	751.535	19.458	
missing	0	0	0	0	0
mode					AUT

Figure 3. Statistical distribution of smaller domain sample gene attributes – GEO dataset

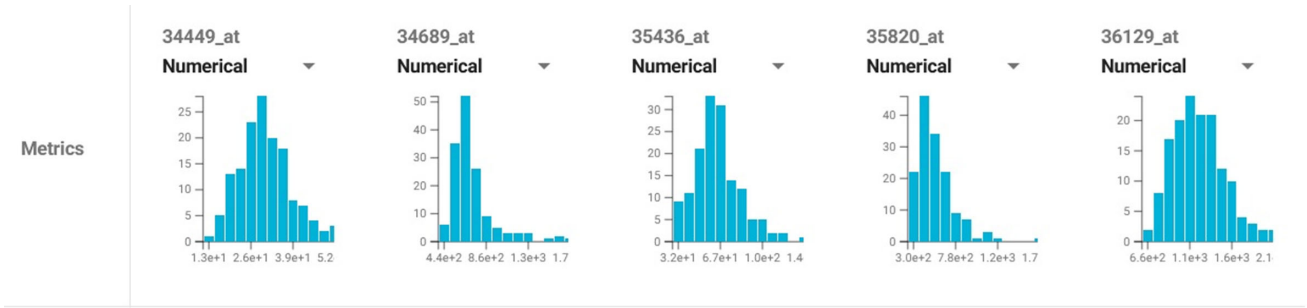
the potential causes for limited classifier performance as reported in the literature. Hence, the statistical features of the dataset were also recorded, and the diverse distribution of values is evident from Figures 3 and 4.

The author performed a comprehensive survey of existing work on the classification of autistic cases from gene expression data. The most recent work on this dataset reported that Support Vector Machine (SVM) classifier [11] yielded the highest performance of ~92% on 124 observations with 101 genes. The previously reported results did not make use of all the instances and, hence, yielded biased results that were inconsistent and susceptible to change based on the observations included in the sample.

Hence, the authors in this work attempted to execute and analyze the performance of the standard traditional, deep, and automated machine learning [19–21] algorithms on the original dataset with a suitable feature ranking algorithm to identify the minimal yet optimal set of genes that could distinguish between autistic and control patients.

This work records the performance of the Naïve Bayesian classifier that yielded an Matthew's correlation coefficient (MCC) of ~70%, reported for the first time in literature on gene expression data for autism, along with an accuracy of ~87%, higher than SVM on feature-reduced data. The previous work did not report on the ranking of classifiers for unbalanced data. Nevertheless, the authors in this work report on the statistical performance measures for unbalanced data as well, and the Naïve Bayes algorithm with the Fast Correlation-Based Feature Selection (FCBF) feature ranking method outperforms the other methods in terms of MCC, balanced accuracy, and accuracy. The methods utilized in this work are described below.

The Orange data mining suite [22, 23] available in Anaconda v3.11 was utilized to train the traditional and deep learning classifiers post-feature selection. Many feature selection



Metrics	34449_at	34689_at	35436_at	35820_at	36129_at
min	12.761	442.140	31.576	302.950	657.799
max	55.605	1805.750	147.169	1840.830	2174.010
nanmean	31.907	771.771	68.969	604.079	1275.975
nanstd	8.143	223.163	19.925	215.760	290.423
median	31.485	722.267	66.660	565.371	1257.240
missing	0	0	0	0	0
mode					

Figure 4. Statistical distribution of larger domain sample gene attributes – GEO dataset

algorithms were implemented on the gene expression data. It was found that FCBF algorithms generated the best combination of features that improved the classifier performance.

2.1. Fast correlation-based feature filtering

Feature selection is a highly regarded step in machine learning as most of the classifiers rely on the goodness of features to raise their performance [24]. The goodness of the features is measured by how well the feature values contribute to discriminating between the classes [25, 26]. In practice, it is essential to retain as good and significant those attributes whose values are not correlated or dependent on other attributes but those that are highly correlated to the target class values [27, 28]. The FCBF is an entropy-based measure that also considers the pairwise correlation between features, thereby eliminating redundancy [29, 30]. This also makes it a suitable and scalable algorithm for large and unbalanced datasets. Predominant correlation based on symmetric uncertainty (SU) is the entropy-based measure that plays the key role in deciding the elimination of a feature from the subset. The SU is based on the entropy of each feature and the information gain of feature pairs [31, 32]. The entropy of individual feature A is calculated as:

$$H(A) = -\sum_{i} P(a_i)\log_2(P(a_i)) \quad (1)$$

The pairwise relationship between any 2 attributes A, B is measured by Equation (2).

$$H(A|B) = -\sum_{j} P(b_j)\sum_{i} P(a_i|b_j)\log_2(P(a_i|b_j)) \quad (2)$$

The information gain is calculated from Equations (1) and (2) as:

$$IG(A|B) = H(A) - H(A|B) \quad (3)$$

From Equation (3), the SU of A and B is defined by:

$$SU(A,B) = 2\left[\frac{IG(A|B)}{H(A)+H(B)}\right] \quad (4)$$

From Equation (4), the predominant correlation heuristic states that a feature "a" is accommodated into the feature subset iff, the SUa,c >¥, where "¥" is a predefined threshold for correlation value for the given dataset. It is also to be noted that there exists no other feature "b" in the dataset such that SUa,b >= SUa,c. Hence, a feature is a predominantly correlated one if it is highly correlated to the class "c" (based on SU) or becomes predominant after the removal of its redundant peers that are correlated by a lower value to the target variable [33–35]. The different feature selection algorithms analyzed in this research are listed below in Figure 5.

The FCBF, Relief-F, information gain, and ANOVA feature ranking methods [35, 36] were applied to the gene expression data. However, the features ranked by the FCBF yielded the highest performance in classification. A combination of hybrid feature selection was also attempted to measure the gene rank across the algorithms. However, the ranking made by the FCBF proved to hold the most contributing gene signature for autism classification. Post-feature ranking and selection, the classifier models were evaluated on both the entire set of features and the subset ranked by the feature selection methods. The entire dataset consumed a lot of time while also yielding poor performance. The choice of methods was based on the scalability of the algorithm and the classifier performance recorded during the literature review as shown in Figure 6.

2.2. Naive Bayes

Classification is faster and more accurate, especially on scalable datasets when dimensionality reduction is done, and the classifier can focus on the significant factors in the dataset that contribute most to

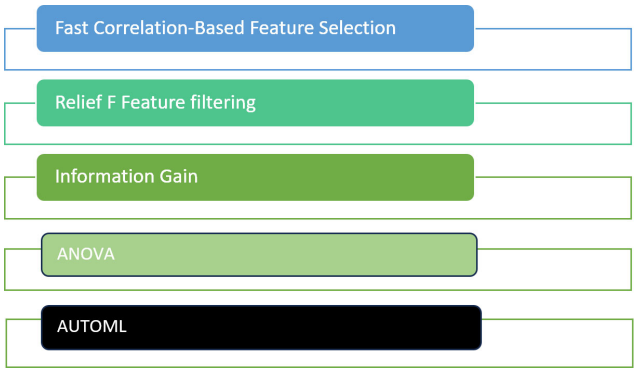


Figure 5. Feature ranking methods for gene signature detection from autism gene expression data

the decision-making process [36, 37]. The Naïve Bayes classifier is a probabilistic machine learning model that is scalable and has proven to yield high accuracy in binary/multinomial classification problems. However, it works on the assumption that the features involved in the dataset are independent of one another and that all attributes contribute equally to determining the target [38, 39]. The Naïve Bayes classifier works primarily based on the Bayes theorem given by:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (5)$$

The probability of X occurring given the certainty that Y is true is calculated from Equation (5). The detailed algorithm description is given in [37, 38].

2.3. Neural network

Neural network classifiers were explored in this paper to establish the performance of network-based classifiers, deep network classifiers, and convolutional neural networks in predicting gene expression data [25]. Neural networks also provide estimates of posterior probability but carry along the advantages of being a nonlinear model-free method. Since this research addresses a binary classification problem, the neural networks here directly decide the membership of an instance based on the following discriminant function.

$$G(a) = P(w_1|a) - P(w_2|a) \quad (6)$$

The record "a" is assigned to class w_1 if the probability $G(a) > 0$ as calculated from Equation (6); otherwise, it is classified as class w_2.

The performance of the classifiers can be enhanced by modifying the bias, number of hidden layers, the computation function at the nodes, and using recurrent and feedforward networks [39, 40]. The neural network classifier was implemented with different parametric values, and as the number of layers increased, the performance of the classifier dwindled. Hence, the neural network was set to computation at just one hidden layer with 100 neurons. The more the number of layers, the more computation time and the lower the performance was recorded on this dataset.

2.4. AutoML (automated machine learning)

The JADBio (Just-Add Data) software suite was employed for experimenting with the AutoML models and evaluating their performance on autism classification from gene expression data [40]. The dataset was loaded, and the basic steps for data formatting were done. The classification models were evaluated based on both the complete set of features and the gene sets selected by the feature selection algorithms. AutoML in JADBio reported the SES (statistically equivalent signature) algorithm to be the most effective in detecting the gene signatures. The SES algorithm also applied the Bayesian networks' constraint-based learning [41]. The SES method discovered latent feature subsets that had high predictive information while also maintaining statistically equivalent performances [42]. The ridge logistic regression and the SVM methods also yielded good classification results. However, the performance was much lower in comparison to the traditional and deep machine learning models.

The results of the different models and the effect of feature signatures on the classifier performance are discussed elaborately in the ensuing section.

3. Experimental Results and Discussions

The results of this research are presented in three sections. The first section emphasizes the effect of feature selection and the gene

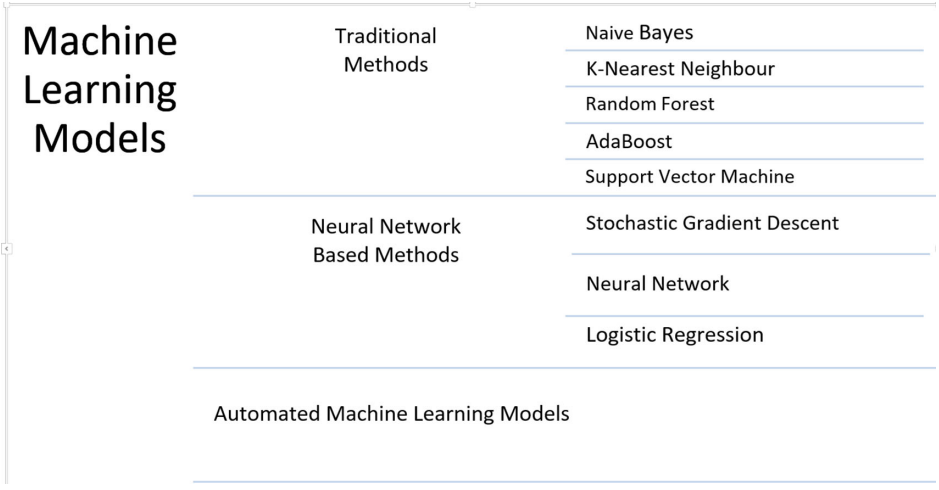


Figure 6. Machine learning models for autism classification

sets that were discovered by the feature selection model. The second section portrays the influence of the gene sets on classifier performance, wherein the performance of the traditional, neural network-based, and Auto ML methods is discussed. The final section compares the results of this work to previous reports on similar data.

3.1. Feature selection

Identifying the most significant genes\gene signatures (combination of genes) that contribute to good discrimination between the class values is one of the most important tasks while building machine learning models for data science applications [15, 21, 36]. In this work, four traditional feature ranking methods and one AutoML feature selection technique were evaluated. The number of features to be ranked was chosen based on the measurable values given by the FCBF algorithm. All the features that had a significance rating of more than 0 were included in the gene subsets.

The gene signatures and their relevant ranking according to the feature selection methods are displayed in Figure 7.

Moreover, it was also supported by the fact that raising or reducing the count of features showed much lower classification performance across the different classifiers. The features not considered significant by FCBF are ranked high by the ANOVA test, Relief-F, and information gain. On proceeding with the features ranked by the different feature selection methods, it was found that the classifier performance was maximum when the FCBF feature subset was input to the classifiers. The number of features was restricted to 22, as beyond that, the filtering value was 0 for all the other genes.

Figure 8 displays the distribution of classes based on the probability values of the pairwise gene combinations taken in the order of their ranking.

The selected genes can distinguish between the classes based on the probability values, thereby signifying the FCBF wrapped in the Naïve Bayes model to be a good choice for autism classification from gene expression data. The gene signatures and their contribution to the respective classes are evident from the distribution of records.

	#	Info. gain	ANOVA	ReliefF	FCBF
1553569 at		0.163	12.287	-0.004	0.122
1556314 a at		0.157	20.333	0.030	0.117
230530 at		0.149	12.954	0.051	0.111
222815 at		0.138	4.882	0.005	0.000
221948 s at		0.136	23.256	0.019	0.100
207084 at		0.135	15.348	0.017	0.000
1555309 a at		0.133	5.423	0.012	0.097
1566690 at		0.128	16.461	0.028	0.000
217055 x at		0.127	8.294	0.017	0.000
91682 at		0.123	16.245	0.015	0.000
34449 at		0.122	15.940	0.028	0.089
215497 s at		0.119	5.577	0.021	0.000
1570102 at		0.119	15.460	0.013	0.000
208819 at		0.119	22.447	0.043	0.000
1557993 at		0.118	4.434	0.016	0.000
220165 at		0.118	10.345	0.032	0.000
209159 s at		0.118	19.167	0.018	0.000
233835 at		0.116	15.962	0.014	0.000

Figure 7. Gene ranking based on the assessed feature ranking methods

Figure 8 shows the distribution of classes when the gene signature pairs in the descending order of rank, generated by the FCBF method, were used to distinguish between the autistic and control instances.

The experimental results derived from Figure 9 indicate the presence of a few outliers, and hence, the author believes that handling outliers in the data prior to feature selection and classification may also contribute to enhancing the classifier accuracy and discovering more accurate gene signatures. Figure 9 also displays the projection of class distribution based on the gene signature generated by FCBF. The classes are clearly distinguishable by the gene signature. The gene signature generated by the AutoML SES feature selection algorithm is displayed in Figure 10.

The contribution of the highly ranked features by the AutoML models generated a very low MCC/accuracy and hence was not considered for further study. Automated methods work on the principle of self-improvement, and hence, it does not provide room for human intervention that minimizes the chances of parametric adjustments, trial and error methods, and hybrid models of implementation.

3.2. Evaluation of classifier performance

Classification is the final step of designating the class of an unknown test dataset. In this work, since the data was unbalanced, it was required to identify appropriate statistical measures to rank the classifiers [34, 42]. Mathew's correlation coefficient and balanced accuracy, along with accuracy, Receiver Operating Characteristic (ROC) analysis, and F-1 score, were measured in this work to establish the classifier performance. This is reported for the first time in the literature. The metrics applied in this research are MCC, Balanced Accuracy (BA), Accuracy (AC) and F1 Score (F1S). They are calculated from Equations (7), (8), (9) and (10) respectively.

$$\epsilon_{\mathbf{MCC}} = \frac{(\mathbf{TP} * \mathbf{TN}) - (\mathbf{FP} * \mathbf{FN})}{\sqrt{(\mathbf{TP} + \mathbf{FN}) * (\mathbf{TN} + \mathbf{FP}) * (\mathbf{TP} + \mathbf{FP}) * (\mathbf{TN} + \mathbf{FN})}} \tag{7}$$

wherein TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively. In this research, the presence of autism is considered positive, while the accurate classification of a control is considered as true negative [8, 9].

$$\epsilon_{\mathbf{BA}} = (\mathbf{Sensitivity} + \mathbf{Specificity})/\mathbf{2} \tag{8}$$

where sensitivity is denoted by the true positive rate and specificity is calculated as the true negative rate. The calculation of the accuracy and F1-score from the TP, FP, TN, and FN values is denoted below.

$$\epsilon_{\mathbf{AC}} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{FP} + \mathbf{TN} + \mathbf{FN}} \tag{9}$$

$$\epsilon_{\mathbf{F1S}} = \frac{\mathbf{2} * \mathbf{TP}}{\mathbf{2} * \mathbf{TP} + \mathbf{FP} + \mathbf{FN}} \tag{10}$$

It is clear from the way the metrics are calculated that the accuracy and F1-score are more biased toward the positive predictions (both true and false) and give lesser weightage to the negative class, thereby causing the results to be biased toward the more populated target value. The evaluation of the models based on the metrics for unbalanced data is a requirement that makes the well-

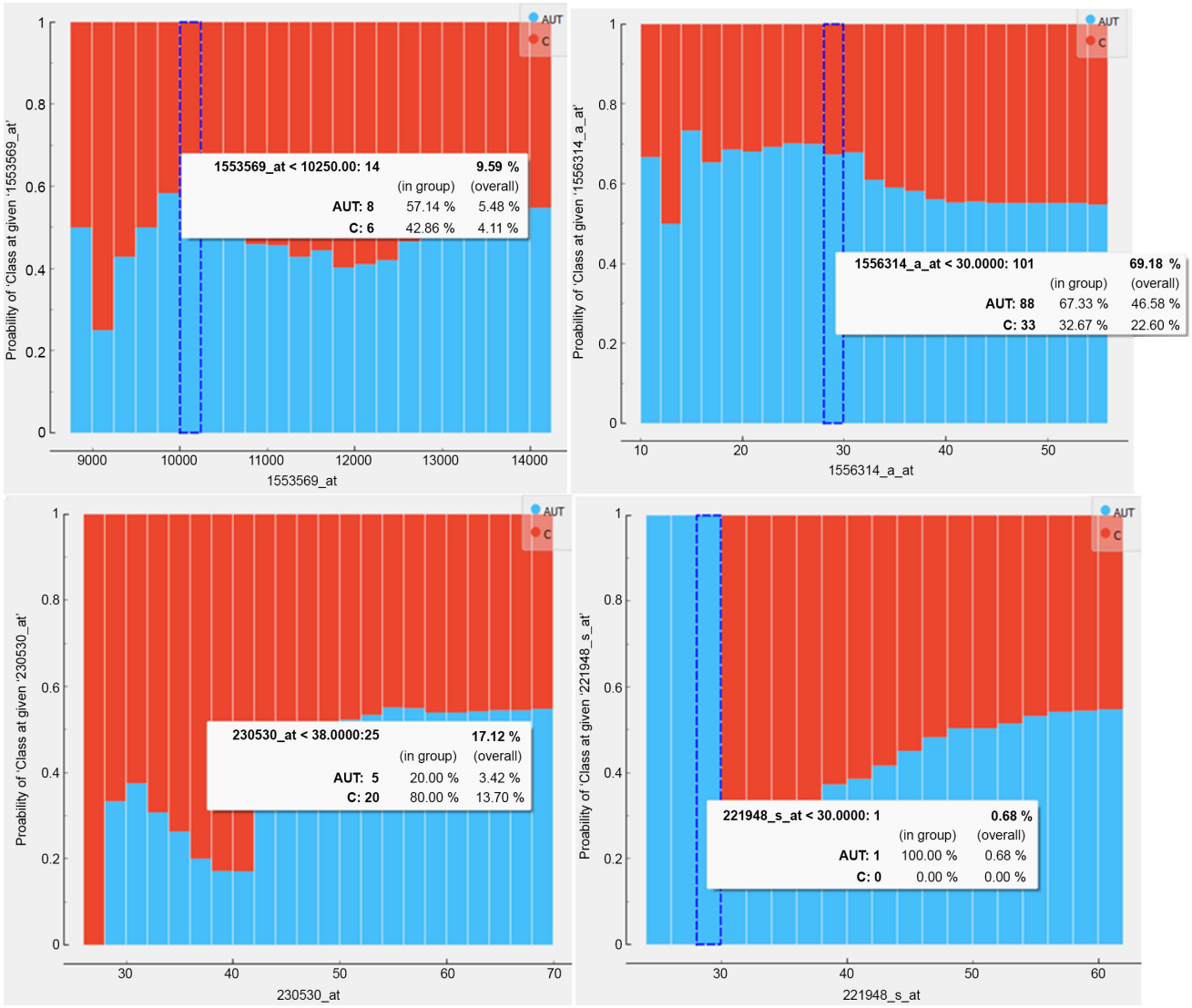


Figure 8. Distribution of classes based on the gene pair combinations of the highly ranked genes

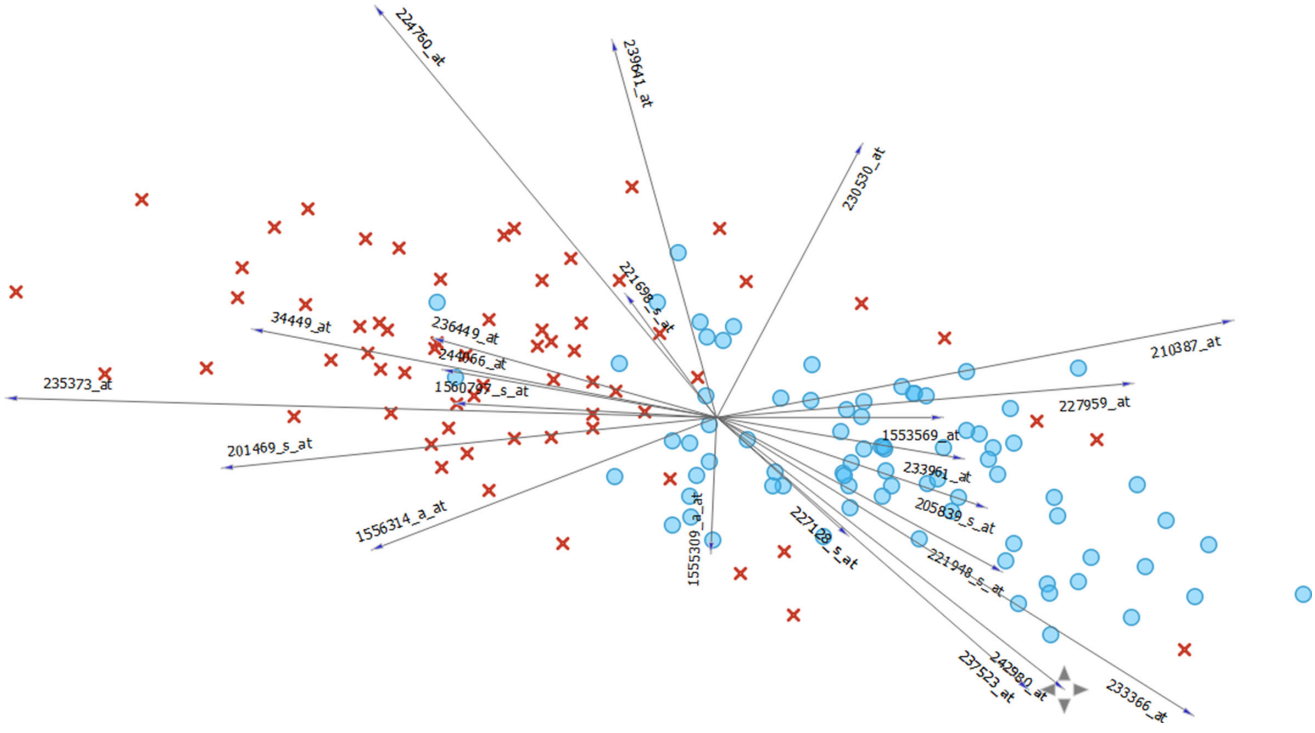


Figure 9. Projection of gene signature and their contribution to autism classification

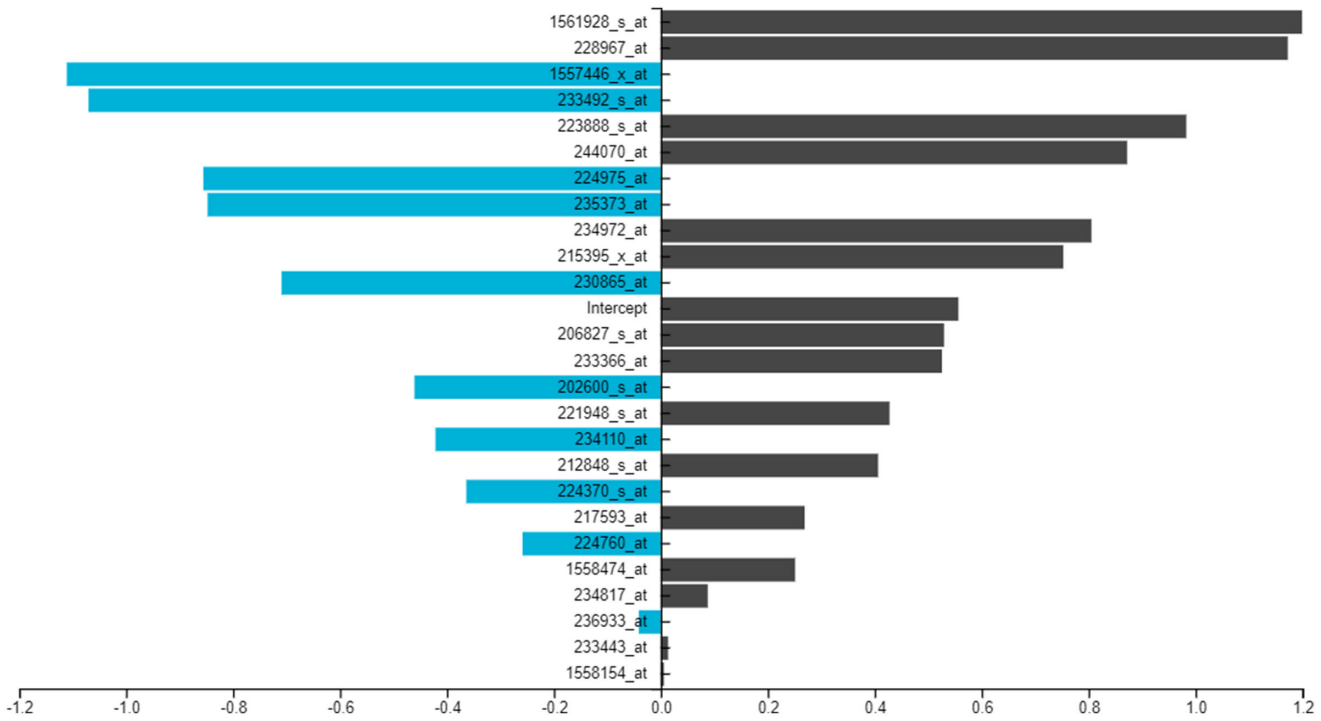


Figure 10. Gene signature for autism classification generated by automated ML models

performing models applicable to augmented gene expression data as well, irrespective of the classes to which the new instances belong. Thereby, the need for sampling and balancing of class values is alleviated even when the machine learning models work on datasets with unbalanced class distribution. The 10-fold cross-validation methods were used across all the learning models [16, 17]. The cumulative results of the different learning models with the features (gene signatures) selected by the FCBF are presented in Table 2.

The ROC curve is a data visualization technique that portrays the relationship between the sensitivity and specificity of a classification model [21, 22, 34]. The curve is plotted for pairs of the TP rate and FP rate for every possible decision threshold of a classification model. The graphical representation of the classifier model performance is shown in Figures 11 and 12.

The gene signature defined by the FCBF model yielded the highest accuracy and MCC with the Naïve Bayes model when compared to the previous work that generated a high accuracy only on an independent test set and the train-test method of evaluation.

Figure 13 exhibits the class distribution as generated by the gene signatures given by the information gain, ANOVA, and Relief-F methods. The gene signatures are not as effective as those generated by the FCBF method (Figure 9) in distinguishing between classes.

3.3. Comparison to existing work

Prior work by Hameed et al. [11] on this gene expression data was reported in 2017. The work involved immense preprocessing of data, and the final evaluation of the classifiers was done based only on the train-test method of evaluation along with an independent test set. Moreover, the authors reported only on the accuracy of the classifiers but did not evaluate the classifiers using statistical

Table 2. Comparative performance of machine learning models on the gene signature generated by FCBF algorithm

S.No	Classifier	MCC	Balanced accuracy (%)	Accuracy (%)	*F*1-score	AUC
	Conventional learning models					
1.	Naïve Bayes	0.74	86.93	86.99	0.88	0.93
2.	K-Nearest Neighbor	0.29	64.55	65.07	0.69	0.7
3.	Random Forest	0.57	78.50	78.77	0.81	0.82
4.	AdaBoost	0.29	64.45	64.38	0.66	0.65
5.	Support Vector Machine	0.65	82.25	82.88	0.85	0.88
6.	CN2	0.36	68.11	67.81	0.69	0.77
	Neural Network Models					
7.	Neural Network	0.67	83.67	83.56	0.85	0.87
8.	Stochastic Gradient Descent	0.41	71.00	69.19	0.72	0.82
9.	Logistic Regression	0.56	78.00	78.23	0.80	0.82
	Automated ML					
10.	Ridge Logistic Regression with SES	0.295	64.7	64.4	0.663	0.68

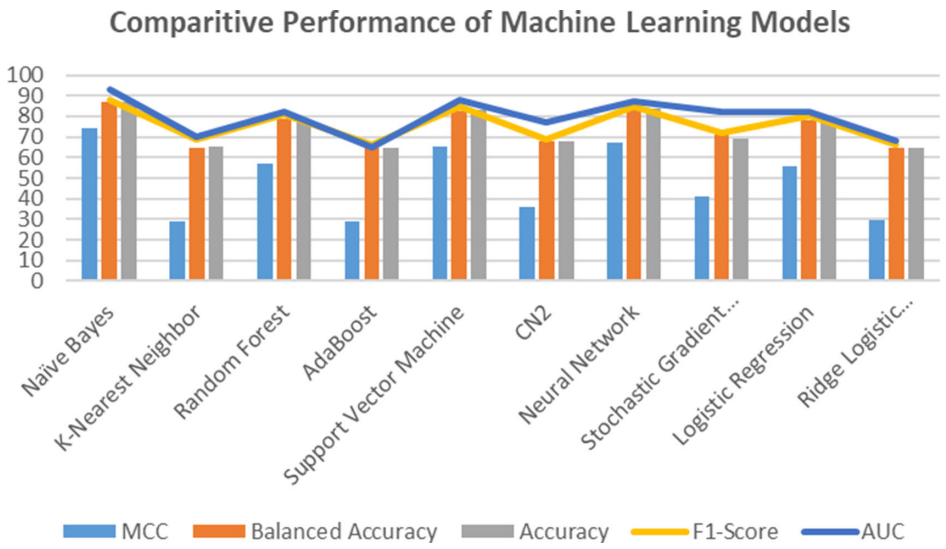


Figure 11. Visualization of classifier performance metrics on gene signature selected by FCBF

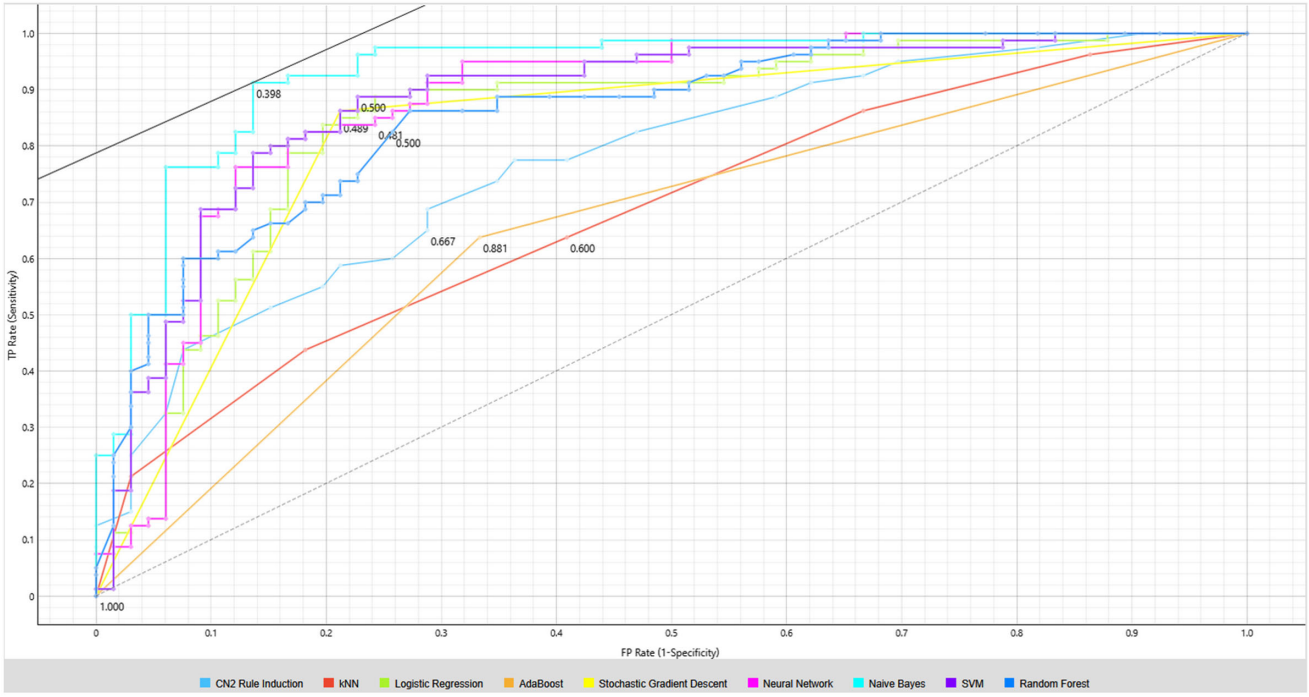


Figure 12. The AUC visualization of the classifier models on autism gene expression data

performance measures for unbalanced datasets. While preprocessing the data, it was also noted that two records were classified as normal in the training set, while the actual data was recorded as autistic in the NCBI database. The previous work implemented only the traditional classifiers with many filtering methods. Automated and deep learning approaches were not considered in the previous work.

The accuracy reported by the previous researchers records a highest of 84.7% on the independent test set when close to 200 genes were involved in the classification [17]. However, in this work, through repeated cross-validation, with a minimal yet optimal gene signature, the Naïve Bayes classifier reports an accuracy of 86.9% along with a balanced accuracy of ~86% and MCC of 0.74, the highest and first reported thus far in literature.

The author, however, believes that analyzing the effect of incremental/decremental feature selection methods on the gene expression data is a definite extension to this work. The possibility of exploring the performance of hybrid feature selection methods and fusion classifiers would be a great step forward in enhancing classifier accuracy on gene expression data. The role of feature construction methods would also open new areas for exploring the possibility of genetic variants in autism and how possible medication/therapy could bring about the constructed feature characteristics in affected individuals. This gene expression data also contains information on the paternal and maternal age of the parents of the individuals involved in the study, both control and autistic. However, there were many missing

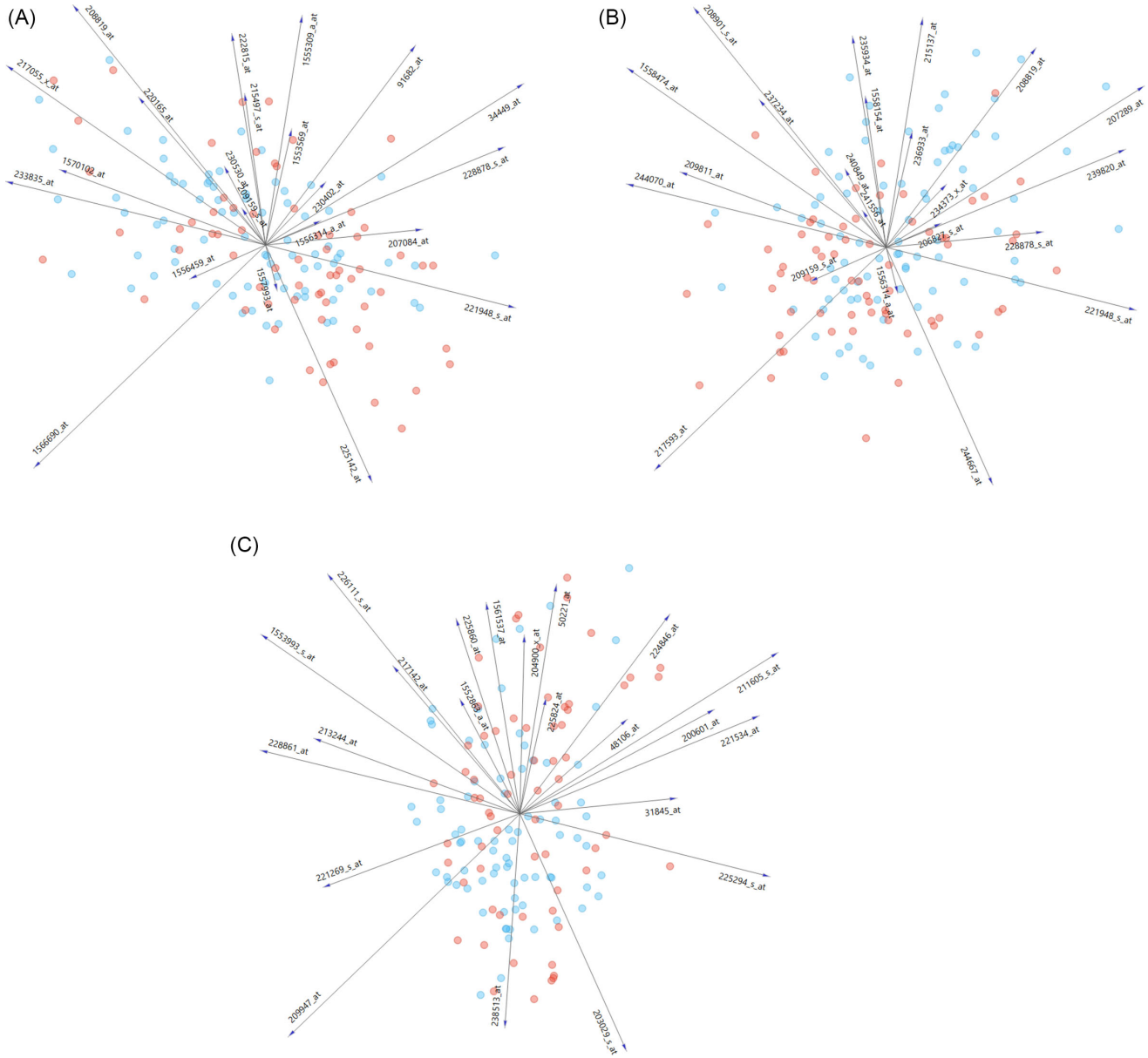


Figure 13. Projection of gene signature and their contribution to autism classification: (A) information gain method, (B) ANOVA, (C) Relief-F

values, and hence, they were not included in this research. A possible way forward would be to identify suitable data cleaning and missing value handling methods in data science and attempt to unearth the effects of maternal/paternal age on triggering autism.

4. Conclusion

Autism is identified as a mental health condition that has been on the rise in recent years across all ethnicities, economic/social/educational backgrounds, parents with normal genetic and medical histories, healthy lifestyles, and safe drug ingestions/absence of complications during pregnancy. Although multiple researchers across diverse backgrounds spanning medical, genetic, and computational fields of study are working hard to identify the possible triggers for this uninvited growth suppression in unsuspecting individuals, there has been very little progress in this area. In vivo assays of gene expression and genetic mutations are a herculean task owing to the intensive number of resources they consume in terms of time, labor, and capital. Computational models have proved to be faster, accurate, and scalable in analyzing and processing large datasets with the advancement of data science and data analytics using machine learning models.

This research has placed focus on the expertise of computational models in acquiring, transforming, processing, analyzing, and evaluating machine learning models to identify potential gene signatures that carry information on autism triggers. This work has evaluated diverse feature ranking and classifier models and has reported on the FCBF with Naïve Bayes classifier to be more accurate in distinguishing between autism and control cases on gene expression data. Analyzing gene expressions is a

noninvasive manner of detecting autism triggers/the presence of autism and hence calls for more research in this sphere for providing advanced care and possible preventive therapy for individuals and parents at greater risk.

Recommendations

The finding revealed that more real-time data and scalable machine learning models would enhance the possibility of automating early autism detection. This would also open avenues for personalized therapy and better survival rates of affected individuals.

Acknowledgment

The author is grateful to the School of Information and Communication Technology, Faculty of ICT and Creative Media, Bahrain Polytechnic, for their administrative support.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by the author.

Conflicts of Interest

Shomona Gracia Jacob is an Associate Editor for *Medinformatics* and was not involved in the editorial review or the decision to publish this article. The author declares that she has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the NCBI database at www.ncbi.nlm.nih.gov.2011 and http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431.

Author Contribution Statement

Shomona Gracia Jacob: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Gogate, A., Kaur, K., Khalil, R., Bashtawi, M., Morris, M. A., Goodspeed, K., . . . , & Chahrour, M. H. (2024). The genetic landscape of autism spectrum disorder in an ancestrally diverse cohort. *NPJ Genomic Medicine*, *9*(1), 72. https://doi.org/10.1038/s41525-024-00458-0

[2] Sheik Abdullah, A., V, K., Geetha, S., & Mishra, U. (2025). Leveraging deep learning for enhanced diagnosis of autism spectrum disorder using resting-state functional magnetic resonance imaging and clinical data. *Results in Engineering*, *25*, 104444. https://doi.org/10.1016/j.rineng.2025.104444

[3] Adako, O., Adeusi, O., & Alaba, P. (2024). Integrating AI tools for enhanced autism education: A comprehensive review. *International Journal of Developmental Disabilities*. Advance online publication. https://doi.org/10.1080/20473869.2024.2392983

[4] Ali, M. T., Gebreil, A., ElNakieb, Y., Elnakib, A., Shalaby, A., Mahmoud, A., . . . , & Elbaz, A. S. (2023). A personalized classification of behavioral severity of autism spectrum disorder using a comprehensive machine learning framework. *Scientific Reports*, *13*(1), 17048. https://doi.org/10.1038/s41598-023-43478-z

[5] Athilakshmi, R., Jacob, S. G., & Rajavel, R. (2023). Automatic detection of biomarker genes through deep learning techniques: A research perspective. *Studies in Informatics and Control*, *32*(2), 51–61. https://doi.org/10.24846/v32i2y202305

[6] Dick, K., Kaczmarek, E., Ducharme, R., Bowie, A. C., Dingwall-Harvey, A. L. J., Howley, H., . . . , & Armour, C. M. (2025). Transformer-based deep learning ensemble framework predicts autism spectrum disorder using health administrative and birth registry data. *Scientific Reports*, *15*(1), 11816. https://doi.org/10.1038/s41598-025-90216-8

[7] Jacob, S. G., Bait Ali Sulaiman, M. M., & Bennet, B. (2023). Feature signature discovery for autism detection: An automated machine learning based feature ranking framework. *Computational Intelligence and Neuroscience*, *2023*(1), 6330002. https://doi.org/10.1155/2023/6330002

[8] Doumari, S. A., Berahmand, K., & Ebadi, M. J. (2023). Early and high-accuracy diagnosis of Parkinson's disease: Outcomes of a new model. *Computational and Mathematical Methods in Medicine*, *2023*(1), 1493676. https://doi.org/10.1155/2023/1493676

[9] Brueggeman, L., Koomar, T., & Michaelson, J. J. (2020). Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Scientific Reports*, *10*(1), 20994. https://doi.org/10.1038/s41598-020-77832-2

[10] Elshoky, B., Ibrahim, O., & Ali, A. (2021). Machine learning techniques based on feature selection for improving autism disease classification. *International Journal of Intelligent Computing and Information Sciences*, *21*(2), 65–81. https://doi.org/10.21608/ijicis.2021.61582.1058

[11] Hameed, S. S., Hassan, R., & Muhammad, F. F. (2017). Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm. *PLoS One*, *12*(11), e0187371. https://doi.org/10.1371/journal.pone.0187371

[12] Gunning, C., Breathnach, Ó., Holloway, J., McTiernan, A., & Malone, B. (2019). A systematic review of peer-mediated interventions for preschool children with autism spectrum disorder in inclusive settings. *Review Journal of Autism and Developmental Disorders*, *6*(1), 40–62. https://doi.org/10.1007/s40489-018-0153-5

[13] Lin, Y., Yerukala Sathipati, S., & Ho, S. Y. (2021). Predicting the risk genes of autism spectrum disorders. *Frontiers in Genetics*, *12*, 665469. https://doi.org/10.3389/fgene.2021.665469

[14] Zaman, N., Ferdus, J., & Sattar, A. (2021). Autism spectrum disorder detection using machine learning approach. In *2021 12th International Conference on Computing Communication and Networking Technologies*, 1–6. https://doi.org/10.1109/ICCCNT51525.2021.9579522

[15] Wu, C., Liaqat, S., Helvaci, H., Cheung, S.-C. S., Chuah, C.-N., Ozonoff, S., & Young, G. (2021). Machine learning based autism spectrum disorder detection from videos. In *2020 IEEE International Conference on E-health Networking, Application & Services*, 1–6. https://doi.org/10.1109/healthcom49281.2021.9398924

[16] Rastegari, M., Salehi, N., & Zare-Mirakabad, F. (2023). Biomarker prediction in autism spectrum disorder using a network-based approach. *BMC Medical Genomics*, *16*(1), 12. https://doi.org/10.1186/s12920-023-01439-5

[17] Singh, N. K., Patel, A., Verma, N., Singh, R. K. B., & Sharma, S. K. (2025). Hybrid deep learning method to identify key

genes in autism spectrum disorder. *Healthcare Technology Letters*, *12*(1), e12104. https://doi.org/10.1049/htl2.12104

[18] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., . . . , & Zupan, B. (2013). Orange: Data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353. https://dl.acm.org/doi/abs/10.5555/2567709.2567736

[19] Farhat, T., Akram, S., Rashid, M., Jaffar, A., Bhatti, S. M., & Iqbal, M. A. (2025). A deep learning-based ensemble for autism spectrum disorder diagnosis using facial images. *PLoS One*, *20*(4), e0321697. https://doi.org/10.1371/journal.pone.0321697

[20] Geetha Ramani, R., & Jacob, S. G. (2013). Prediction of cancer rescue p53 mutants *in silico* using Naïve Bayes learning methodology. *Protein and Peptide Letters*, *20*(11), 1280–1291. https://doi.org/10.2174/09298665113209990046

[21] Geetha Ramani, R., & Jacob, S. G. (2013). Prediction of P53 mutants (multiple sites) transcriptional activity based on structural (2D&3D) properties. *PLoS One*, *8*(2), e55401. https://doi.org/10.1371/journal.pone.0055401

[22] Demšar, J., & Zupan, B. (2024). Hands-on training about data clustering with orange data mining toolbox. *PLOS Computational Biology*, *20*(12), e1012574. https://doi.org/10.1371/journal.pcbi.1012574

[23] Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. (2019). Applications of supervised machine learning in autism spectrum disorder research: A review. *Review Journal of Autism and Developmental Disorders*, *6*(2), 128–146. https://doi.org/10.1007/S40489-019-00158-X

[24] Luo, L. (2021). Research on image classification algorithm based on convolutional neural network. *Journal of Physics: Conference Series*, *2083*(3), 032054. https://doi.org/10.1088/1742-6596/2083/3/032054

[25] Mehralizadeh, B., Soleiman, P., Nikkhoo, S., Rahimi, M., Kargar, A., Masoumi, F., & Moradi, H. (2023). Multi-modal ASD screening system: A preliminary study. In *2023 11th RSI International Conference on Robotics and Mechatronics*, 228–234. https://doi.org/10.1109/ICRoM60803.2023.10412541

[26] Meneses do Rêgo, A. C., & Araújo-Filho, I. (2024). Leveraging artificial intelligence to enhance the quality of life for patients with autism spectrum disorder: A comprehensive review. *European Journal of Clinical Medicine*, *5*(5), 28–38. https://doi.org/10.24018/clinicmed.2024.5.5.350

[27] Mohammed, M. B., Salsabil, L., Shahriar, M., Tanaaz, S. S., & Fahmin, A. (2021). Identification of autism spectrum disorder through feature selection-based machine learning. In *2021 24th International Conference on Computer and Information Technology*, 1–6. https://doi.org/10.1109/ICCIT54785.2021.9689805

[28] Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific Reports*, *10*(1), 19747. https://doi.org/10.1038/s41598-020-76635-9

[29] Wankhede, N., Kale, M., Shukla, M., Nathiya, D., Roopashree, R., Kaur, P., . . . , & Koppula, S. (2024). Leveraging AI for the diagnosis and treatment of autism spectrum disorder: Current trends and future prospects. *Asian Journal of Psychiatry*, *101*, 104241. https://doi.org/10.1016/j.ajp.2024.104241

[30] Asha, P., Raj, A., Ramani, B. V., Suresh, L. P., Mary, S. P., & Posonia, A. M. (2024). Implementation of various AI techniques for autism spectrum disorder detection. In *2024 7th International Conference on Circuit Power and Computing Technologies, 1*, 117–121. https://doi.org/10.1109/ICCPCT61902.2024.10673050

[31] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, 994–1004. https://doi.org/10.1016/j.procs.2020.03.399

[32] Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., . . . , & Sherry, S. T. (2022). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

[33] Serajian, M., Marini, S., Alanko, J. N., Noyes, N. R., Prosperi, M., & Boucher, C. (2024). Scalable *de novo* classification of antibiotic resistance of *Mycobacterium tuberculosis*. *Bioinformatics*, *40*(Supplement_1), i39–i47. https://doi.org/10.1093/bioinformatics/btae243

[34] Shafiq, M., Ali, A., Ali, F., & Choi, J.-G. (2024). Machine learning, data mining, and IoT applications in smart and sustainable networks. *Sustainability*, *16*(18), 8059. https://doi.org/10.3390/su16188059.

[35] Webb, G. I. (2010). Naïve Bayes. In C. Samuel, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 713–714). Springer. https://doi.org/10.1007/978-0-387-30164-8_576

[36] Yang, F.-J. (2018). An implementation of Naive Bayes classifier. In *2018 International Conference on Computational Science and Computational Intelligence*, 301–306. https://doi.org/10.1109/CSCI46756.2018.00065

[37] Xin, M., & Wang, Y. (2019). Research on image classification model based on deep convolution neural network. *EURASIP Journal on Image and Video Processing*, *2019*(1), 40. https://doi.org/10.1186/s13640-019-0417-8

[38] Zhou, X., Du, H., Xue, S., & Ma, Z. (2024). Recent advances in data mining and machine learning for enhanced building energy management. *Energy*, *307*, 132636. https://doi.org/10.1016/j.energy.2024.132636

[39] Yuan, H., Yu, K., Xie, F., Liu, M., & Sun, S. (2024). Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare. *Medicine Advances*, *2*(3), 205–237. https://doi.org/10.1002/med4.75

[40] Khaire, U. M., & Dhanalakshmi, R. (2019). Optimizing feature selection parameters using Statistically Equivalent Signature (SES) algorithm. In *2019 4th International Conference on Information Systems and Computer Networks*, 625–629. https://doi.org/10.1109/ISCON47742.2019.9036211

[41] Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(4), 451–462. https://doi.org/10.1109/5326.897072

[42] Zaffar, M., Hashmani, M. A., & Savita, K. S. (2019). Comparing the performance of FCBF, Chi-Square and Relief-F filter feature selection algorithms in educational data mining. In *Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology*, 151–160. https://doi.org/10.1007/978-3-319-99007-1_15

How to Cite: Jacob, S. G. (2025). Gene Signatures for Autism Classification: Mining Biological Markers for Autism from Gene Expression Data. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52024698