Journal of Ambient Intelligence and Humanized Computing (2021) 12:3257–3270
https://doi.org/10.1007/s12652-020-02155-z

ORIGINAL RESEARCH

Predicting autism spectrum disorder from associative genetic markers of phenotypic groups using machine learning

Karthik Sekaran[1] · M. Sudha[1]

Received: 27 February 2020 / Accepted: 25 May 2020 / Published online: 31 May 2020

Abstract
Machine learning is a discipline of artificial intelligence, geared towards the development of various critical applications. Due to its high precision, it is widely adopted in the process of extracting useful hidden patterns and valuable insights from complex data structures. Data extracted from the real-time environment might contain some irrelevant information. The presence of noise in the data degrades the model performance. Gene expression is an important source, carries the genetic information of species. Gene expression pattern reveals the significant relationship between genes associated with several diseases. But due to irregular molecular interactions and reactions occurs during the transcription process, the gene expressions are minimally affected. It causes a detrimental effect on the identification of biological markers of the diseases. To address this problem, a novel gene selection strategy is proposed to identify the candidate gene biomarkers from the genomic data. Signal to Noise ratio with logistic sigmoid function, Hilbert–Schmidt Independence Criterion Lasso, and regularized genetic algorithm amalgamation finds the optimal features. The proposed system is tested with the microarray gene expression dataset of autism spectrum disorder (ASD), accessed from gene expression omnibus repository. FAM104B, CCNDBP1, H1F0, ZER1 are identified as the candidate biomarkers of ASD. The methodical performance evaluation of the proposed model is examined with widely used machine learning algorithms. The proposed methodology enhanced the prediction rate of ASD and attained an accuracy of 97.62%, outperformed existing methods. Also, this system could act as a significant tool to assist the medical practitioners for accurate ASD diagnosis.

Keywords Autism spectrum disorder · Biomarkers · Dimensionality reduction · Hilbert–schmidt independence criterion lasso · Machine learning · Regularized genetic algorithm

1 Introduction

Autism spectrum disorder (ADS) is a type of neurodevelopmental illness, also be identified as pervasive developmental disorder (PDD) (Faras et al. 2010), predominantly affects the children's of age group between 2 and 5. The prevalence of ASD is increased by 1 in every 68 in the US (Oztan et al. 2018). The functional impairment of ASD is high when compared with other neurodevelopmental and psychiatric disorders (Leyfer et al. 2006). Most of the children with ASD are not diagnosed properly until the age of 4 years (Oh et al. 2017). Genetic and environmental factors influence the cause of developing ASD (Hallmayer et al. 2011). Recent studies revealed that ASD can also be inherited from their parents through genetic contact. It cannot be spontaneously developed but substantially it could rise with the impact of the growth on the victim. Children with ASD suffer from learning disabilities, impaired verbal and non-verbal communication, weakened social interactions, behavioral overlaps, impulsivity, lack of focus on regular work, repetitive actions, improper eye contact, etc. (Hallmayer et al. 2011; Duda et al. 2016). The presence of ASD in an individual can be detected through its preliminary symptoms. Diagnosing ASD in earlier stages with proper treatment increases the chance of progressive recovery, otherwise, it becomes ineffective. Gene expression profiles of autistic and normal individuals vary in their genetic patterns. The discrimination factor between both the categories is helpful in the identification of potential gene biomarkers. In this paper,

✉ M. Sudha
msudha@vit.ac.in

Karthik Sekaran
skarthik@vit.ac.in

[1] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

gene expression profiles of autistic and normal children are analyzed with sophisticated computational methodologies.

Signal to noise ratio (SNR) is an effective technique to calculate the strength of signals. In this case, the gene features are mapped as signals (Gunavathi and Premalatha 2015). Initially, SNR is calculated for all the gene features present in the dataset. The logistic sigmoid function is applied to select the gene features concerning its threshold frequency that determines the minimum signal strength. The non-linear relationship between the selected features is computed with Hilbert–Schmidt Independence Criterion Lasso (HSIC-Lasso) technique. The candidate gene subset selection process is performed with regularized genetic algorithm (RGA). In this algorithm, the fitness function is penalized by the ridge regression (L2) technique to guide the population towards an optimal solution in the solution space. The given dataset is finally reduced with the identified gene subset. Supervised machine learning algorithms are employed to train the data for the classification of samples. Extreme gradient boosting (XGBoost), support vector machines (SVM), logistic regression (LR), multi-layered perceptron neural network (MLP-NN), and bayes net (BN) algorithms undertake the process of classifying the samples. These models are validated by Leave-one-out cross validation technique and the performance of these models is evaluated using standard metrics.

The rest of the paper is organized as follows. In Sect. 2, existing literature related to this experiment is analyzed in-depth. The problems identified from the previous works are discussed in Sect. 3. Section 4 describes the workflow of the proposed system and the inferences made from the identified genes are given. In Sect. 5, result analysis is performed with standard evaluation metrics and the results are plotted in the graph. Section 6 describes the highlights of the proposed model and discusses the significance of this work. Section 7 concludes the work and further improvements of the system are discussed.

2 Background Study

After the successful completion of the Human Genome Project in 2003, comprehensive research was conducted by the expertise from heterogeneous scientific communities to analyze protein structures, the interaction between genotype, phenotypes, and identification of genetic patterns associated with diseases (Collins et al. 2003). The advent of high-performance computing technologies simplifies the process of handling high dimensional genomic data. But still, there is always room for development in the way of discovering diagnostic biomarkers of diseases. Sophisticated computational algorithms work effectively on solving the problem of handling large volumes and different varieties of data. Targeted therapy finds more significant improvement in patient's condition on a disease compared with other treatment procedures. In a recent study, the prognostic markers of breast cancer are identified through computational models (McKenna et al. 2018; Kolch and Fey 2017). Progressive improvement on a patient with stable responsiveness towards the treatment is achieved. Similarly, for many complex neurological diseases, categorized as neurodevelopmental, neurodegenerative, and psychiatric disorders gene therapy tends to provide better response over other treatment procedures such as anti-depressants, chemical imbalance inhibitors, and so on (Stevens et al. 2019; Gök 2019). Studies on how the genetic factors influence an individual's mental ability, reasoning, and processing could be more helpful in developing personalized drugs for curing the ailment (Sekaran and Sudha 2020; Karthik and Sudha 2020).

A 21 gene-model was derived to classify schizophrenic samples from the control group. The dataset is collected from gene expression omnibus (GEO) and its accession number is GSE17612 and GSE21935. SVM-RFE is used for feature selection. Five different supervised machine learning algorithms k-NN, RF, Extra Trees, AdaBoost, and SVM are used for classification. Parameters of each algorithm are tuned for better performance (Logotheti et al. 2016).

Feature selection based on mutual information is performed on colon and lymphoma gene expression datasets for optimal gene selection. For classification, SVM with different kernels, ANN, and k-NN is used. Leave-one-out cross validation technique (LOOCV) is used for evaluating the model (Vanitha et al. 2015). Fuzzy quick reduct algorithm is used based on similarity measure with customization for selecting informative gene features. The dimension of the dataset is reduced using information gain (IG). Leukemia, lung, and ovarian datasets are used to conduct this experiment. RF algorithm is used for classifying the data (Arunkumar and Ramakrishnan 2018).

A hybrid genetic algorithm and learning automata (GALA) algorithm is proposed to classify six different cancer datasets. The time complexity of this method is O(G.m.n3) (Motieghader et al. 2017). A decision tree (DT) based classification model is developed to predict Alzheimer's disease. The dataset is obtained from GEO and its accession number is GDS810. It contains 31 samples (9 controls and 22 AD) with a total number of 22,283 genes. Out of them, 69 genes are differentially expressed, considered as informative features (Kumar and Singh 2018).

Ant colony optimization technique (ACO) and cellular learning automata model (CLA) is developed for gene feature selection. To learn the complicated relationship of data, CLA is utilized. K-NN, Naïve Bayes (NB), and SVM classifiers are used to evaluate the performance of the models on four different datasets (Sharbaf et al. 2016).

Optimal feature selection using a t-test is performed on the heart disease dataset (GDS4527, 3690) from GEO. A k-fold CV with 4 folds is estimated with NB, SVM, and AdaBoost classifiers. Based on the t-value, 18 informative genes are selected for the classification process (Neelima and Prasad Babu 2017). Iterative transductive-SVM (IT-SVM) technique is developed to minimize the overall risk to solve the transductive problem. Leukemia and colon dataset is used to experiment. For feature selection, the SVM-RFE technique is adopted. The results of IT-SVM are benchmarked with TSVM and progressive T-SVM (Tajari and Beigy 2012).

Factor analysis and feature score criterion (FA-FSC) method, a hybrid concept is developed to identify top informative genes. Leukemia and Breast cancer datasets are used to conduct this experiment. The SVM classifier is adopted for classification. The best result is obtained with 12 top informative genes extracted from the dataset (Gour et al. 2011). The entropy filtering based feature selection technique is applied to three different gene expression datasets. A fuzzy soft set similarity-based classifier is developed for classification. The results are compared with the fuzzy k-NN classifier (Kalaiselvi and Inbarani 2013). A novel hybrid algorithm (GA + SVM) is developed for feature identification. Also, two other feature selection techniques called information gain (IA) and RF-based model is benchmarked with GA + SVM. Six different ML algorithms namely NB, C4.5 DT, RF, k-NN, SVM with linear and Gaussian kernel is used for classification. Top 20 informative genes are extracted using the GA + SVM technique (Scheubert et al. 2012).

Particle swarm optimization algorithm (PSO) with DT combined hybrid classifier is developed to evaluate 11 different gene expression datasets. Also, the performance of the model is compared with Self-organizing map algorithm (SOM), SVM, back-propagation neural networks (BPNN), NB, classification and regression tree (CART) and artificial immune recognition system algorithm (AIRS) (Chen et al. 2014). Recursive feature addition (RFA) is proposed for selecting gene biomarkers. Also, a new algorithm namely the lagging prediction peephole optimization technique (LPPO) is developed for classification. These algorithms are evaluated using six benchmarked gene expression datasets. A model is compared with SVM-RFE and leave-one-out calculation and sequential forward selection (Liu et al. 2011).

A new decision support system, called application specific intelligent computing (ASIC) which adopts an intelligent method to select optimal features as parameters for any learning algorithm to make a good prediction from the clinical data. The genetic algorithm with a new relative fitness function is used to find optimal parameters. BPNN algorithm performed the classification of clinical data (Sudha 2017). An ensemble technique for feature selection with RFE and bayes based error filter (BBF) is developed for selecting informative genes. To perform classification, SVM is adopted. Leukemia dataset is taken for experimenting (Bennet et al. 2015).

The thromboembolic syndrome prediction system is developed using auto-encoder (AE). The performance of AE is compared with principal component analysis (PCA) and AE shows better results in terms of reducing the size and dimensions of the dataset than PCA. Also, the same model is benchmarked with 11 other gene expression datasets (Khalili et al. 2016). A neural network-based gene identification technique is developed to classify Alzheimer's at its early stage. Along with NN, SVM, IG ratio, and Gini coefficient are benchmarked for its performance comparison (Barati and Ebrahimi 2016).

Empirical Bayes technique is used to select features from the breast cancer dataset. The accession number of datasets is GSE1456 and GSE20711. The number of features from the dataset is reduced from 12,000 to 398 genes (López-González and Dávila 2017). A hybrid Discrete Wavelet Transform (DWT) and Moving Window Technique (MWT) is developed for gene feature selection. K-fold CV is performed for evaluating the model. A combination of NB, SVM, and k-NN is used as a classifier (Bennet et al. 2014).

An mRMR combined with the artificial bee colony algorithm (ABC) is proposed for the best gene feature selection. SVM is used to evaluate the performance of the model based on the classification rate. Six different datasets are used to analyze the robustness of the model. Also, mRMR-PSO and mRMR-GA are developed and compared where mRMR-ABC shows better results than the other models (Alshamlan et al. 2015). A meta-heuristic search optimization algorithm called Binary Bat Algorithm (BBA) is proposed for improving the selection of optimal features from data. This algorithm is designed based on the behavior of microbats (Nanda and Panda 2014). An ensemble learning model is developed for diabetes classification with a large scale imbalanced dataset. G-mean and F-score are used for performance evaluation of the classifier (Wei et al. 2017).

3 Problem formulation

Data with higher dimensionality decreases the efficiency of computational models on extracting useful insights. Redundant data can make the model capture the noise, results in the poor generalization of samples. The complexity of the system increases and deteriorates the effective utilization of computational resources. Another issue is the sample size of the gene expression profiles. Most of the genomic data have a very less number of samples compared with their gene features.

Let x ($\in \mathbb{R}^d$) and y ($\in \mathbb{R}$) defines the domain of input vector x and output values y respectively. The original dataset of the system is represented as,

$$x = \left[x_1, .., x_n\right] \in \mathbb{R}^{d*n}$$

$$y = [y_1, .., y_n]^T \in \mathbb{R}^n$$

where ^{T} represents the transpose of the predictor variable. In genomic data, the number of clinical samples is very less with the number of gene features. The algorithm aims to find the gene features were, $(m < d)$ of the given input vector x to predict the output y.

1. The learning models tend to over-fit, as the number of training samples becomes very less.
2. The selection of diagnostic biomarkers becomes tedious with more redundant features.
3. Computational complexity will be more.

4 Materials and methods

The proposed framework follows three steps on finding potential genetic markers of ASD (1) Informative gene selection, (2) candidate subset identification, (3) model training, and evaluation. The architecture diagram and workflow of the proposed system is represented as Figs. 1 and 2.

4.1 Data acquisition

The dataset used to conduct this experiment is accessed from the GEO database; stores curated gene expression of various diseases, managed by the National Center for Biotechnology Information (NCBI) (Edgar et al. 2002).

Agilent 44 K Human whole-genome array G4112F, GPL6480 (Kuwano et al. 2011) microarray is used to obtain oligonucleotide gene expressions from the blood samples. The accession number of the datasets used in this experiment is GSE26415. Table 1 shows a brief description of the dataset.

The dataset contains two different groups namely autistic and control. Each group has 21 samples each. In total, 42 samples with 19,194 gene expression probes are presented in the dataset.

4.2 Proposed biomarker selection mechanism

The identification of significant features improves the overall model performance. Selecting the feasible feature selection technique is the key to sort out the best features. Generic feature selection methods such as filter, wrapper, embedded and hybrid techniques provide robust algorithms for feature selection (Chandrashekar and Sahin 2014). The filter method applies statistical methods on identifying the best feature based on the scoring factor and relationship between the variables (Hameed et al. 2017). Wrapper technique adopts learning models such as support vector machines, logistic regression, etc. that binds the process of selecting the best feature based on the performance of these evaluators. During every iteration, the worst features are eliminated, and the best subset of features that guarantees high accuracy is selected. But, the problem in this method lies in its computational inefficiency due to poor search strategy.

The embedded method fuses the qualities of the filter and wrapper method to select more optimal features. In most cases, regularization models that come under this category are highly successful when compared with previous methods. The hybrid technique combines two or more techniques to make a new model that suits well for different scenarios.

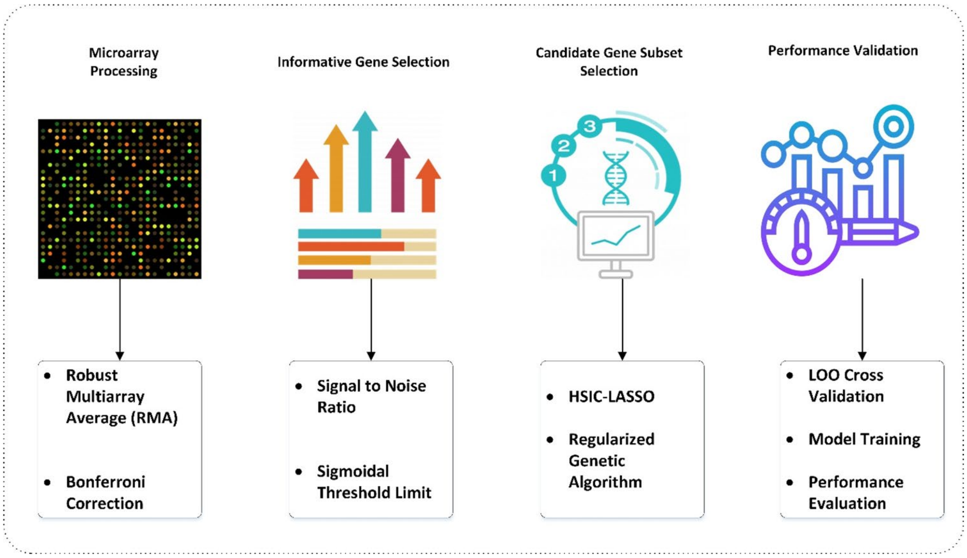

Fig. 1 Architecture diagram of proposed work model

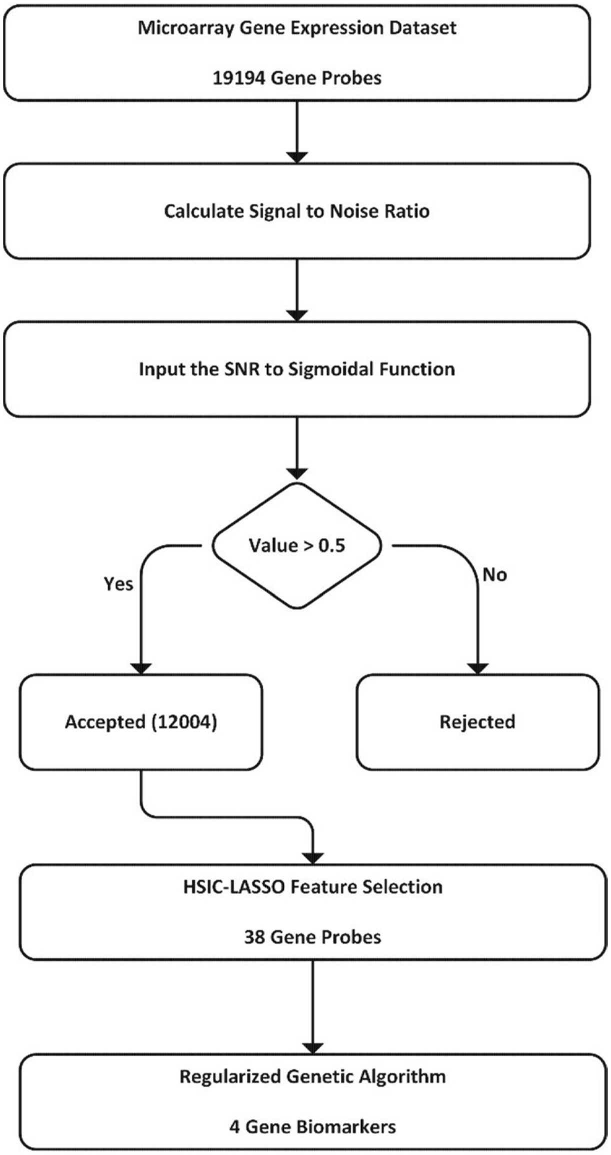
Microarray Gene Expression Dataset
19194 Gene Probes
Calculate Signal to Noise Ratio
Input the SNR to Sigmoidal Function
Value > 0.5
Yes
No
Accepted (12004)
Rejected
HSIC-LASSO Feature Selection
38 Gene Probes
Regularized Genetic Algorithm
4 Gene Biomarkers

Fig. 2 Workflow of tSNR-HSIC-Lasso-RGA

Table 1 Dataset description

Details	Source information
Data Repository	Gene expression Omnibus
Accession Number	GSE26415
Disease type	Autism disorder
Platform	Agilent-014850–4 × 44 K
Type of data	Gene expression
Number of samples	42
Number of features	19,194
Case (disease)	21 samples
Control	21 samples
Data type	Numeric
Class	(0—control, 1—case)

Other than these techniques, many algorithms are developed, inspired by biological processes. Genetic algorithm, particle swarm optimization, bat algorithm, gravitational search algorithm, wolf search algorithm, etc. are some popular algorithms, provides a solution to optimization problems.

4.2.1 Signal to noise ratio

Signal to noise ratio (SNR) is a measure used to compare the desired signal to the background noise. It can determine the strength of the signal and takes decibel as a unit for measurement. The ratio between the power of a signal and the power of noise calculates SNR of an input signal. An alternative statement of SNR is defined as the reciprocal of the coefficient of variation. It is represented as the ratio of mean to standard deviation of any measurement or a signal. In our case, the input signal is mapped with the gene features, where the strength of the gene can be calculated with SNR. From Eq. (1) SNR is calculated for each gene feature.

$$SNR = \frac{\mu}{\sigma} \tag{1}$$

4.2.2 Thresholded signal to noise ratio

The strength of the gene features is measured with SNR in the previous phase. But the limit to define the minimum strength for selecting the best features is not fixed. In tSNR, the logistic sigmoid function is adapted to sort out the features from its standard threshold point 0.5. The SNR value of each gene feature is inputted to the logistic sigmoidal function. If the output from the function satisfies the threshold criteria, then that gene is selected for the next phase, otherwise eliminated. In Eq. (2), the logistic function is defined with its parameters.

$$tSNR = \frac{L}{1+e^{-k}} \tag{2}$$

- e = logarithm base.
- L = maximum value of the curve (L = 1).
- k = steepness of the curve.

Table 2 Parameters of genetic algorithm

Parameters	Values
Population	50
Chromosome type	Binary 0, 1
Number of generations	20
Mutation probability	0.02
Crossover probability	0.8
Elitism count	2

4.2.3 Least absolute shrinkage and selection operator (Lasso)

Lasso is a regression analysis technique that finds the linear dependency of the feature vectors based on the relationship between the input feature and output value (Yamada et al. 2014). It works well in the process of feature selection and regularization (L1) to select a more optimal feature subset. The general form of Lasso is given in Eq. (3).

$$\min_{\alpha \in \mathbb{R}^d} \frac{1}{2} ||y - X^T \alpha||_2^2 + \lambda ||\alpha||_1 \tag{3}$$

Here, α is the coefficient vector, the kth term of the regression coefficient is denoted by α_k, $l1$ and $l2$ norms are represented as $||.||_1$, $||.||_2$ and λ is the regularization parameter. The main limitation of Lasso is in its inability to find the non-linearity in features.

4.2.4 Hilbert–Schmidt independence criterion Lasso

The identification of the non-linear relationship between high dimensional data is complex and computationally expensive. HSIC-Lasso finds the non-redundant features with a strong dependency on the output value (Climente-González et al. 2019). The significant part of HSIC-Lasso lies is in its kernel representation. Gaussian kernel transforms the feature vectors and mapping it into a non-linear representation (Nandagopal et al. 2019). HSIC-Lasso is computed from the given Eq. (4).

$$\min_{\alpha \in \mathbb{R}^d} \frac{1}{2} ||\overline{L} - \textstyle\sum_{k=1}^{d} \alpha_k \overline{K}^{(k)}||_{Frob}^2 + \lambda ||\alpha||_1 \tag{4}$$

$||.||_{Frob}$ is the Frobenius norm, centered gram matrices are represented as $\overline{K}^{(k)}$ and $\overline{L}$. The kernel functions of the equation are $K(x, x')$ and $L(y, y')$ The interpretation of Eq. (4) is rewritten with an expansion as Eq. (5).

$$\begin{aligned} \frac{1}{2} ||\overline{L} - \textstyle\sum_{k=1}^{d} \alpha_k \overline{K}^{(k)}||_{Frob}^2 &= \frac{1}{2} HSIC(y, y) \\ - \textstyle\sum_{k=1}^{d} \alpha_k HSIC(u_k, y) &+ \frac{1}{2} \textstyle\sum_{k,l=1}^{d} \alpha_k \alpha_l HSIC(u_k, uy) \end{aligned} \tag{5}$$

4.2.5 Genetic algorithm

In 1960, John Holland introduced the concept of the genetic algorithm (GA), inspired by Darwin's evolution theory. Later, in 1989, David E. Goldberg, a student of Holland, expanded his work on GA (Goldberg and Holland 1988). This metaheuristic algorithm is developed based on the natural selection of the evolutionary process of species. GA is used to generate the best solutions to solve real-time problems through optimization and search strategies. It functionally depends on three important biological operations such as selection, crossover, and mutation. This method is one of the highly successful optimization strategies, provides effective solutions for many cases. The aim of applying GA in gene expression profiles is to identify the best final solution that reveals significant biomarkers of ASD. The search space is confined to the minimal level as HSIC-Lasso reduced the dimension of the data from 12,004 features to 38. Moreover, it simplifies the complex operation of generating solutions from the population. The job of the objective function is to maximize the fitness value as the fitness function defined in this problem statement determines the best solution based on the accuracy of a learning model. The parameters of the GA are given in Table 2.

4.2.5.1 Initial population and chromosome representation A search space in GA has all feasible solutions at each point. The goal of this algorithm is to find the best solution other from all. This process is also called as tracing the "global optimum" on convergence. The initial population of the genetic algorithm generated from the gene expression data contains the genetic features combined and formed as chromosomes. Each chromosome in the population is represented as bits in the form of binary strings 0 and 1. 0 indicates the feature is not considered and 1 indicates that the feature is selected for a particular chromosome. In such a way, all the chromosomes in the initial population are randomly generated in the search space. Exploitation and Exploration are the two important functions of an evolutionary algorithm (Eiben and Schippers 1998). Finding a good solution in the search space is achieved through the crossover, called exploitation. Exploration, a process of searching for all the solutions in search space is achieved through mutation. Also, these genetic operators help the algorithm to converge towards the optimal point, otherwise, global optima, when the parameters are selected wisely (Srinivas and Patnaik 1994).

4.2.5.2 Fitness function A fitness function in an optimization problem guides the parameters to move towards the optimal solution. It summarizes the design of the model and evaluates whether the model achieves the optimal solution or not. In general, the fitness function can be defined as either maximization or minimization function. The fitness function in this GA is selected as a maximization function, which evaluates the fitness of a chromosome from the accuracy of the learning model. Random forest (RF), an ensemble learning model calculates the fitness for the given solution X. The fitness function of the GA is given in Eq. (6).

$$FitnessFunctionF = X - \alpha \times \frac{S1}{S} \tag{6}$$

where F represents the fitness calculated and X is the accuracy score evaluated by a learning model for the given solution. Here, α is a control parameter, used to adjust the weight ranges from [0–1]. S1 is the number of features selected in the current fitness and S is the total number of features. The accuracy of the model calculated from the features selected in current fitness can be calculated from Eq. (7).

$$Acc = \frac{t}{n} \tag{7}$$

In the above equation, t denotes the number of correctly classified instances and n is the total number of instances of the dataset.

4.2.5.3 Genetic operators Crossover and Mutation adopt single-point crossover and bit-flip mutation techniques for performing the genetic operations respectively. From the theoretical studies and empirical results conducted on various applications by applying GA, the best values for crossover and mutation are given in the range between 0.6 to 0.95 and 0.001 to 0.05 respectively (Sharma et al. 2015).

However, when the values are very less, it might lead to poor convergence. At the same time, if the values are high, then the algorithm could not be able to find the global optimum. Deciding the best parameters leads to better convergence towards the solution point.

4.2.5.4 Termination condition Every algorithm must have termination criteria to stop the process by moving towards infinite iterations. The following are addressed as valid conditions to stop the execution of GA.

1. A fixed number of generations are reached.
2. The expected minimum solution is achieved.
3. Allocated execution time is exceeded.
4. Producing a solution that is no longer optimized further.
5. Manual Interruption

4.2.6 Regularization

Tikhonov Regularization also called Ridge regression or (L2 regularization) is a regression-based penalizing technique. This method solves the problem of overfitting by penalizing the coefficients of the features with higher coefficient values (Wang et al. 2013). A penalty term λ is added to the loss or cost function, which is a squared magnitude of the coefficients. By the concept of "shrinking" in L2, the redundant features are eliminated from the feature set. In Eq. (8) the cost function is defined as root mean squared error (RMSE). The error rate is calculated through RMSE.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_0(x)^{(i)} - y^{(i)})^2 \tag{8}$$

The penalty term of L2 regularization technique is given as Eq. (9)

$$R = \lambda \sum_{j=1}^{n} \theta_j^2 \tag{9}$$

4.2.7 Proposed regularized genetic algorithm

In normal cases, the fitness of the previous solution is not considered while calculating the fitness for the current one. So, the final solution might not be the best, as it misses the optimal point of convergence from its previous population. Through regularization, the learning model that calculates the fitness of the solution is penalized to generate a better solution based on the fitness value of the previous solution. In such a way, this model guides the solution towards the optimal point of convergence. The complete L2 regularization form is represented in Eq. (10)

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_0(x)^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \tag{10}$$

Algorithm: tSNR-HSIC-LASSO-RGA

Input: Population Size (*p*), Mutation Rate (*m*), Crossover (*c*), Elitism Count (*ec*), Elitism Rate, (*er*), Number of Iterations (*i*), Fitness Function (*f*), Penalizing Term (*pt*), Solution (S)

Output: Final Population with the best Solution

```
generate initial population $p$ with feasible solutions
for j = 1 to $i$ do
    choose $ec = er$. p
    select the best $ec$ solutions and map them to $p1$
    $c = (p - ec)/2$
    for k = 1 to $c$ do
        select two random solutions $S_A$ and $S_B$ from $p$
        generate new $p_2$ as $S_C$ and $S_D$ by single-point crossover to $S_A$ and $S_B$
    end for
    for $l$ in 1 to $c$ do
        select a random solution $S_n$ from $p_2$
        perform mutation on each bit of $S_n$ to $m$ and generate $S_n$'
        if $S_n$' is unfeasible towards the optimal solution
            update $S_n$' with a feasible solution
        end if
        update $S_n$ with $S_n$' in $p_2$
        evaluate the fitness $f$ of $S_n$' and store it in $f'$
        if $(f' > f)$
            $f = f + pt$
        re-evaluate fitness $f$ of solution $S_n$' and update $f'$
    end for
    update = $p1+p2$
end for
update the best solution $S$ in $p$
```

The proposed algorithm identifies four gene subsets as optimal features from the given gene expression samples. Those genes are represented in Table 3 with the description.

4.3 Heat map visualization

A heat map visualizes the gene expression data in a clustered grid form. The hierarchical clustering technique is applied to the candidate gene subset to plot this heat map. Each row represents individual samples and the column indicates the gene features. The color palette in the heat map represents the changes occurred in the gene expression. Heat maps combine with different clustering methods to group the samples based on their similarity pattern to identify the regulation of genes.

Table 3 Identified biomarkers of ASD

Biomarkers	Description
FAM104B	Family with sequence similarity 104 member B
CCNDBP1	Cyclin D1 binding protein 1
H1F0	H1 histone family member 0
ZER1	zyg-11 related cell cycle

In Fig. 3, the up-regulated and down-regulated genes are represented with red and green color respectively. Black color indicates the absence of regulation.

4.4 Gene pathway analysis

The similarity between the functional genes is identified through gene network and pathway analysis. A large number of genes functional association data supports the tool to create the network with logical interactions.

There are many gene regulation prediction tools are available such as STRING, FunCoup, VisANT, GeneMania, etc. Figure 4 is generated with the help of GeneMania (Warde-Farley et al. 2010). It provides a more user-friendly environment, flexible, faster response, and gives accurate results when compared with other tools. The input for this tool is a list of genes. It generates a significant set of genes that are strongly interrelated with the input genes. The nodes highlighted with stripes in Fig. 3 are the biomarkers of ASD and other nodes are correlated genes.

The importance of analyzing the genetic interactions is to find the pattern and its correlated functionalities associated

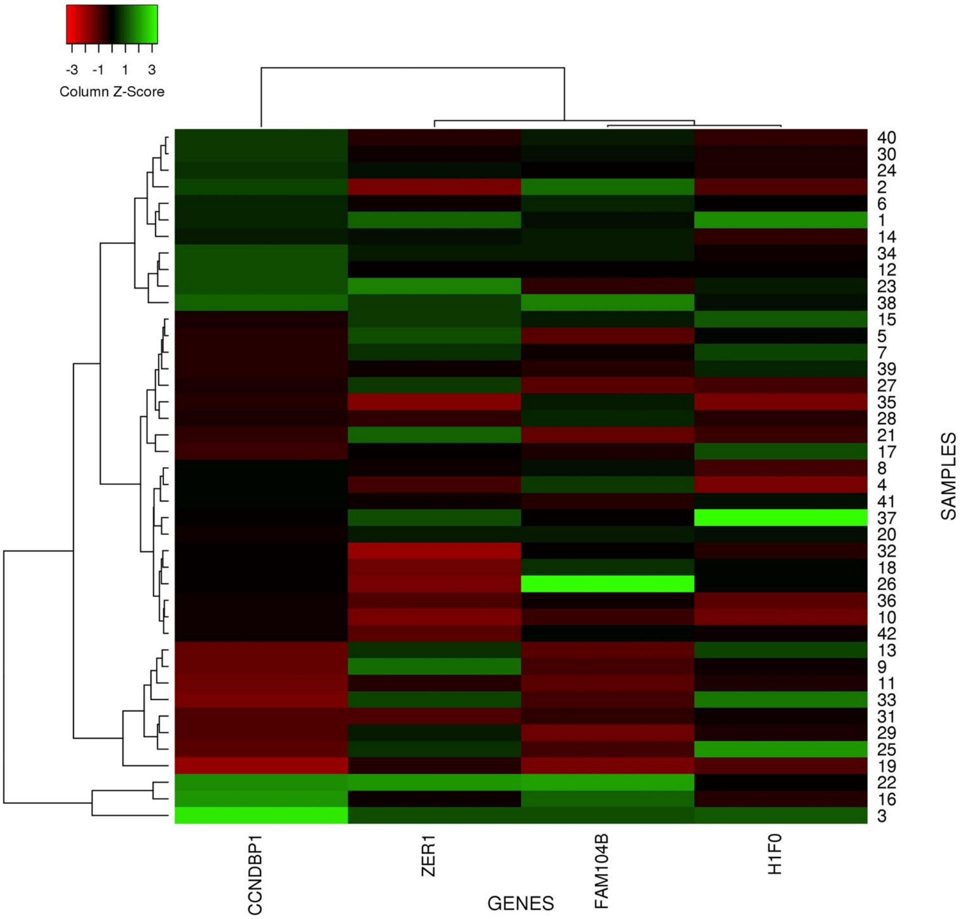

Fig. 3 Heat map representation of identified genes

with any disease. In Fig. 4, each node represents a gene, and the network represents the relationship between the genes.

Co-expression, colocalization, and genetic interaction between the genes in Fig. 4 are represented with violet, grey, and green colored stripes respectively. From the graph, FAM104B strongly related to FAM104A, CCNDBP1 binds with BGPM, ZNF627, MAN1B1, ZER1 closely associated with TCEB2, CUL2 and DFFB. The gene marker H1F0 interrelated with more significant genes such as HIST1H1C, TBC1D22A, DFFB, etc. More advanced studies on these identified genes could improve the diagnosis of disease more accurately. Also, it provides a way to find the correlation between the diseases that are strongly associated with an autism spectrum disorder.

4.5 Statistical inferences on candidate markers

A statistical evaluation is conducted on the identified gene biomarkers of ASD to verify its significance. Limma library is used to calculate the statistical parameters for the gene expression (Ritchie et al. 2015). Bonferroni correction is made on the data to adjust and correct the p-value. The result analysis is given in Table 4.

The significance level 0.05 is used to identify genes that are differentially expressed through a t-test. Log2 transformation is performed to identify the regulation of genes. From the analysis FAM104B, CCNDBP1 are identified as upregulated and H1F0, ZER1 are down-regulated, as the positive and negative LogFC scores indicate up and down-regulation. Moreover, the adjusted p-values are very low that represents the selected biomarkers that are differentially expressed.

5 Experimental results

The proposed model is executed on the system runs on Windows 10 Operating System with Intel Core i7 (6th Generation), 2.6 GHz, and 8 GB RAM. The experimental results show that there is a significant impact on the results achieved by the learning model after applying the tSNR-HSIC-Lasso-RGA algorithm on the gene expression dataset.

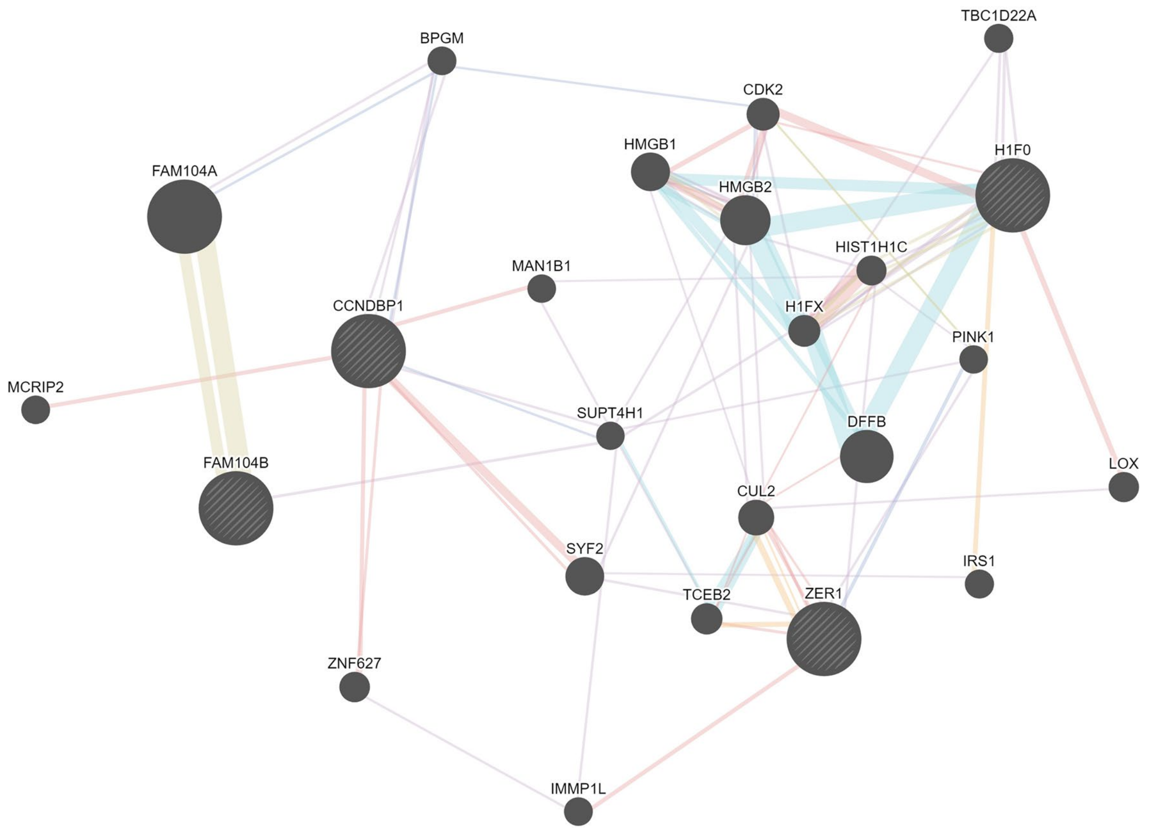

Fig. 4 Network representation of gene interactions

Table 4 Statistical inferences

ID	Adj.p-value	T	LogFC
A_23_P11341	0.0013	5.03	– 0.58
A_23_P26243	0.0053	– 3.65	– 0.40
A_23_P29257	0.0234	2.84	0.40
A_24_P170826	0.0244	2.82	0.30

5.1 Supervised learning algorithms

Five standard machine learning algorithms were employed to train and classify the samples of the given dataset with the identified biomarkers. Bayes net (BN), logistic regression (LR), support vector machine (SVM), multilayered perceptron neural network (MLP-NN), and extreme gradient boosting (XGBoost) algorithms evaluate the performance of the system with selected features.

5.2 Model Performance evaluation

The ML algorithms are validated through a leave-one-out cross-validation technique. The performance of the learning models is calculated with standard evaluation metrics. Confusion matrix is an important metric used to evaluate the performance of classification models. It has four parameters such as true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). These parameters identify the correctly and incorrectly classified instances from the samples given to test the performance of the model.

Sensitivity is the measurement of the proportion of actual positives with correctly identified instances with actual positives. Specificity is the calculation of the proportion of actual negatives with correctly identified instances (Karthik et al. 2018). The probability of subjects with a positive screening test, which have the disease is called positive predicted value (PPV). As an inverse, the probability of subjects with a negative screening test, which does not have the disease, is called a negative predicted value (NPV).

Out of all the models, the best result is obtained with the combination of the tSNR-HSIC-Lasso-RGA-XGBoost model. In Table 5, the results of the model are projected with its evaluation metrics.

5.3 Comparison of feature selection techniques

The performance of the proposed work model is benchmarked with three well-performing feature selection

Table 5 Performance of XGBoost on proposed FS method

Metrics	Formula	Results (%)
Accuracy	(TP + TN)/(TP + TN + FP + FN)	97.62
Sensitivity	TP/(TP + FN)	100
Specificity	TN/(TN + FP)	95.45
Precision(Pr)	TP/(TP + FP)	95.23
Recall(Re)	TP/(TP + FN)	100
P. P. V	TP/(TP + FP)	95.24
N. P. V	TN/(TN + FN)	100

techniques. Correlation-based feature selection (filter method), recursive feature elimination (wrapper), and particle swarm optimization (meta-heuristic optimization technique) are the existing algorithms used to compare the performance with the proposed algorithm. The number of genes selected from these methods is given in Table 6.

These methods are formally evaluated with the same ML Algorithms. The achieved result is benchmarked with the results of existing models, given in Table 7.

The results obtained from benchmarked feature selection methods are plotted in Figs. 5, 6, 7, and 8 where each graph represents the accuracy, sensitivity, specificity, and root mean squared error of the models respectively. The results projected in the graph highlights the significance of the proposed work.

Table 6 Number of features selected by each FS methods

Feature selection technique	Number of features
Correlation-based feature selection	32
Recursive feature elimination	31
Particle swarm optimization	15
Minimum redundancy maximum relevance	9
tSNR-HSIC-Lasso-RGA	4

Table 7 Comparison of results between existing and proposed system

Results	Acc (%)	Sen (%)	Spec (%)	PPV (%)	NPV (%)
SVM (Oh et al. 2017)	93.8	100	87.5	88.9	100.0
k-NN (Oh et al. 2017)	93.8	100	87.5	88.9	100.0
LDA (Oh et al. 2017)	68.8	62.5	75.0	71.4	66.7
Proposed	97.6	100	95.45	100	95.45

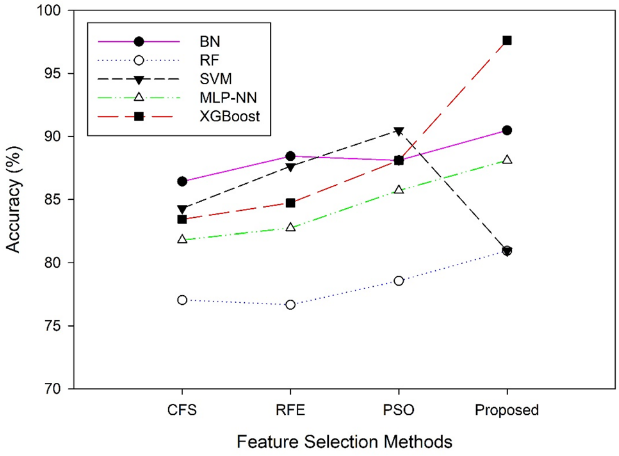

Fig. 5 Accuracy obtained from feature selection methods with five classifiers

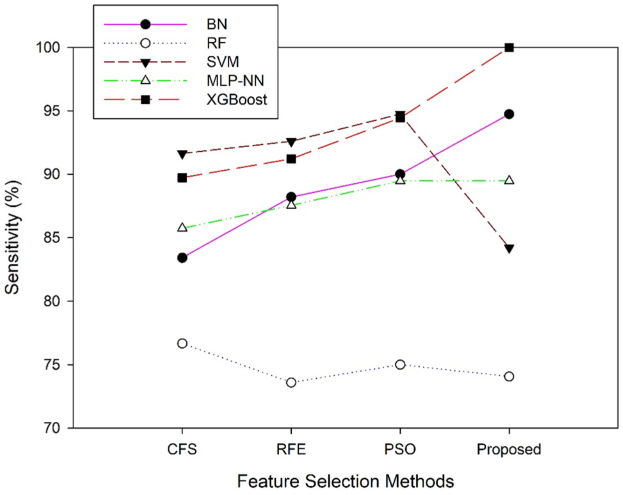

Fig. 6 Sensitivity obtained from feature selection methods with five classifiers

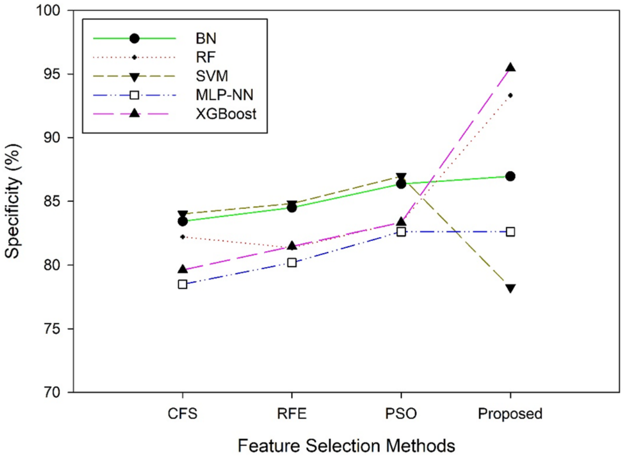

Fig. 7 Specificity obtained from feature selection methods with five classifiers

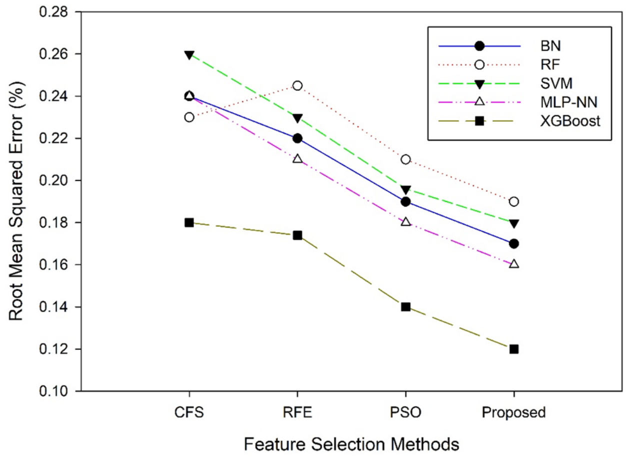

Fig. 8 RMSE obtained from feature selection methods with five classifiers

5.4 Performance of the proposed model on Alzheimer's disease dataset

To validate the performance of the model proposed to identify the biomarkers of ASD, the same methodology is applied to Alzheimer's gene expression profiles. The dataset accession number is GSE1297. The number of samples is 31, out of which 9 are from control and the remaining 22 are Alzheimer's affected. The proposed model finds 24 novel gene biomarkers of Alzheimer's disease. The performance is further examined with the machine learning algorithms. The results are given in Table 8.

Based on the observation from the above table, the SVM model discriminates the samples well over other models with 92.95% accuracy. SVM performs better than XGBoost in Alzheimer's sample discrimination. But the optimal results are yet to be achieved with the support of more sophisticated mathematical models.

6 Discussions

This experimental work intends to identify the candidate gene biomarkers of ASD to discriminate the samples based on the underlying genetic pattern and its variations. Signal to noise ratio finds the strength of the gene features and the logistic threshold function filters out the strong features out from the bag contains all the gene features. Initially, out of 19,194 features, 12,004 are selected for the next stage. HSIC-Lasso algorithm tremendously reduced the number of genes from 12,004 to 38 by finding the non-linear relationship between the features. Candidate gene biomarkers are obtained through Regularized Genetic Algorithm. At last, four genes are selected as an optimal subset for predicting ASD. The extreme Gradient Boosting algorithm reaches the best accuracy rate of 97.62%, outperformed other benchmarked algorithms. The things achieved from this experiment and listed below.

Table 8 Performance of the machine learning classifiers on the biomarkers of Alzheimer's disease

Model	Acc (%)	Sen (%)	Spec (%)
RF	90.48	94.74	86.96
SVM	92.95	94.07	92.33
LR	80.95	84.21	78.26
BN	80. 95	78.26	84.21
MLP	85.71	89.47	82.61
XGBoost	91.62	93.80	91.45

1. Optimal gene biomarkers are identified.
2. The dimension of the data is significantly reduced.
3. The accuracy of the model is improved.

Similar studies with gene expressions for various neurological disorders improve the understanding of brain studies. It will be helpful to identify the influential genetic factors on the growth of complex disorders.

7 Conclusion and future enhancements

The proposed HSIC-Lasso-RGA model achieved better results on identifying ASD associated genetic markers from the gene expression profiles. Four potential genes are found to be the transcriptomic biomarker signatures of ASD. This model discriminates the autistic and control samples effectively. By selecting more optimal genes, the computational complexity of the model is reduced. Moreover, these biomarkers are considered as robust markers to diagnose the presence of ASD.

Gene expression profiling is an important procedure in identifying the genetic cause of any disease. Applying more sophisticated computational models in this field can help locate the genes that are responsible for the progression of diseases. In the field of psychiatric and neurodevelopmental research, the frequency of producing data is minimal than other diseases. It brings down the chance of exploring more valuable insights from complex disorders. More data helps to develop robust computational models with better predictability rate. In the future, generating heterogeneous data could increase the chance of discovering new biological genetic patterns for the disorders having a high genetic risk factor of developing the disease. Rapid advancements in this field can bring the possibilities of "Personalized Medicine" closer to reality.

8 Supplementary Materials

The supplementary source files are available at https://www.github.com/karthiksekaran/Autism-Biomarker-Discovery.

References

Alshamlan H, Badr G, Alohali Y (2015) MRMR-Abc: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Biomed Res Int 2015:604910. https://doi.org/10.1155/2015/604910

Arunkumar C, Ramakrishnan S (2018) Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data. Future Comput Inform J 3(1):131–142

Barati M, Ebrahimi M (2016) Identification of genes involved in the early stages of Alzheimer disease using a neural network algorithm. Gene Cell Tissue 3(3):e38415. https://doi.org/10.17795/gct-38415.

Bennet J, Arul Ganaprakasam C, Arputharaj K (2014) A discrete wavelet based feature extraction and hybrid classification technique for microarray data analysis. Sci World J 2014:195470. https://doi.org/10.1155/2014/195470

Bennet J, Ganaprakasam C, Kumar N (2015) A hybrid approach for gene selection and classification using support vector machine. Int Arab J Inf Technol (IAJIT) 12:695–700

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

Chen K-H, Wang K-J, Tsai M-L, Wang K-M, Adrian AM, Cheng W-C, Yang T-S, Teng N-C, Tan K-P, Chang K-S (2014) Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform 15(1):49

Climente-González H, Azencott C-A, Kaski S, Yamada M (2019) Block Hsic Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics 35(14):i427–i435

Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. Science 300(5617):286–290

Duda M, Ma R, Haber N, Wall DP (2016) Use of machine learning for behavioral distinction of autism and adhd. Transl Psychiatry 6(2):e732

Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30(1):207–210

Eiben AE, Schippers CA (1998) On evolutionary exploration and exploitation. Fundam Inform 35(1–4):35–50

Faras H, Ateeqi NA, Tidmarsh L (2010) Autism spectrum disorders. Ann Saudi Med 30(4):295–300

Gök M (2019) A novel machine learning model to predict autism spectrum disorders risk gene. Neural Comput Appl 31(10):6711–6717

Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. Mach Learn 3(2):95–99

Gour DK, Jain YK, Pandey GS (2011) The classification of cancer gene using hybrid method of machine learning. Int J Adv Res Comput Sci 2(2)

Gunavathi C, Premalatha K (2015) Cuckoo search optimisation for feature selection in cancer classification: a new approach. Int J Data Min Bioinform 13(3):248–265

Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J et al (2011) Genetic heritability and shared environmental factors among twin pairs with autism. Arch Gen Psychiatry 68(11):1095–1102

Hameed SS, Hassan R, Muhammad FF (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

Kalaiselvi N, Inbarani HH (2013) Fuzzy soft set based classification for gene expression data. arXiv Preprint arXiv:1301.1502

Karthik S, Perumal RS, Mouli PC (2018) Breast cancer classification using deep neural networks. In: Knowledge computing and its applications. Springer, pp 227–241

Karthik S, Sudha M (2020) Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network. In: Evolutionary intelligence. Springer, pp 1–16

Khalili M, Majd HA, Khodakarim S, Ahadi B, Hamidpour M (2016) Prediction of the thromboembolic syndrome: an application of artificial neural networks in gene expression data analysis. J Paramed Sci 7(2):15–22

Kolch W, Fey D (2017) Personalized computational models as biomarkers. J Pers Med 7(3):9

Kumar A, Singh TR (2018) Computational mining of genomic and proteomic data to gain insight for Alzheimer's disease (Ad)

Kuwano Y, Kamio Y, Kawai T, Katsuura S, Inada N, Takaki A, Rokutan K (2011) Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children. PLoS ONE 6(9):e24723

Leyfer OT, Folstein SE, Bacalman S, Davis NO, Dinh E, Morgan J, Tager-Flusberg H, Lainhart JE (2006) Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. J Autism Dev Disord 36(7):849–861

Liu Q, Sung AH, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y (2011) Gene selection and classification for cancer microarray data based on machine learning and similarity measures. BMC Genom 12(5):S1

Logotheti M, Pilalis E, Venizelos N, Kolisis F, Chatziioannou A (2016) Studying microarray gene expression data of schizophrenic patients for derivation of a diagnostic signature through the aid of machine learning. Biometr Biostat Int J 4(5):00106

López-González K, Dávila C (2017) Predicting survivability using breast cancer subtype with transcriptomic profiles. In: IIE annual conference. Proceedings. Institute of Industrial; Systems Engineers (IISE), pp 1406–1411

McKenna MT, Weis JA, Brock A, Quaranta V, Yankeelov TE (2018) Precision medicine with imprecise therapy: computational modeling for chemotherapy in breast cancer. Transl Oncol 11(3):732–742

Motieghader H, Najafi A, Sadeghi B, Masoudi-Nejad A (2017) A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Inform Med Unlocked 9:246–254

Nanda SJ, Panda G (2014) A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm Evol Comput 16:1–18

Nandagopal V, Geeitha S, Vinoth Kumar K, Anbarasi J (2019) Feasible analysis of gene expression—a computational based classification for breast cancer. Measurement 140:120–125

Neelima E, Prasad Babu MS (2017) Optimizing genome features using T-test to classify the gene expressions as coronary artery disease prone and salubrious. J Theor Appl Inf Technol 95(16)

Oh DH, Kim IB, Kim SH, Ahn DH (2017) Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. Clin Psychopharmacol Neurosci 15(1):47

Oztan O, Jackson LP, Libove RA, Sumiyoshi RD, Phillips JM, Garner JP, Hardan AY, Parker KJ (2018) Biomarker discovery for disease status and symptom severity in children with autism. Psychoneuroendocrinology 89:39–45

Ritchie ME, Phipson B, Di Wu, Yifang Hu, Law CW, Shi W, Smyth GK (2015) Limma powers differential expression analyses for Rna-sequencing and microarray studies. Nucleic Acids Res 43(7):e47–e47

Scheubert L, Luštrek M, Schmidt R, Repsilber D, Fuellen G (2012) Tissue-based Alzheimer gene expression markers-comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. BMC Bioinform 13(1):266

Sekaran K, Sudha M (2020) Predicting drug responsiveness with deep learning from the effects on gene expression of obsessive-compulsive disorder affected cases. Comput Commun 151:386–394

Sharbaf FV, Mosafer S, Moattar MH (2016) A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics 107(6):231–238

Sharma N, Anpalagan A, Obaidat MS (2015) Evolutionary algorithms for wireless network resource allocation. In: Modeling and simulation of computer networks and systems. Elsevier, pp. 629–52

Srinivas M, Patnaik LM (1994) Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans Syst Man Cybern 24(4):656–667

Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E (2019) Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. Int J Med Inform 129:29–36

Sudha M (2017) Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice. J Med Syst 41(11):178

Tajari H, Beigy H (2012) Gene expression based classification using iterative transductive support vector machine. Int J Mach Learn Comput 2(1):76

Vanitha CD, Arockia DD, Venkatesulu M (2015) Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Comput Sci 47:13–21

Wang F, Chawla S, Liu W (2013) Tikhonov or Lasso regularization: which is better and when. In: 2013 IEEE 25th international conference on tools with artificial intelligence. IEEE, pp. 795–802

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M et al (2010) The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38(suppl_2):W214–W220

Wei X, Jiang F, Wei F, Zhang J, Liao W, Cheng S (2017) An ensemble model for diabetes diagnosis in large-scale and imbalanced dataset. In: Proceedings of the computing frontiers conference. ACM, pp. 71–78

Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M (2014) High-dimensional feature selection by feature-wise kernelized Lasso. Neural Comput 26(1):185–207

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.