Autism Spectrum Disorder Risk Gene Prediction Using Improved Salp Swarm Algorithm and Enhanced Convolution Neural Network Algorithm

Dr.J. Anitha[1*]

Abstract

A group of Neuro developmental disorders is Autism Spectrum Disorder (ASD) which is characterized by communication skills and social interaction difficulties. In past few years, there is a rapid increase in ASD and its root symptom's cause is not yet determined. Comparatively large effort is needed in the existing system using Bayes network and there is no universally accepted technique to construct a network from data. An Enhanced Convolution Neural Network (ECNN) and Improved Salp Swarm Algorithm (ISSA) is proposed to solve these issues and for effective ASD classification. The task corresponds to classifying a long non-coding RNA (lncRNA) gene would cause a disease or not. After class balancing, discretization is applied for converting continuous values into discrete values and for optimal gene selection, ISSA algorithm is used. From genomic data, candidate gene biomarkers are identified using this gene selection. Every possible feature subset is computed for minimizing irrelevant features and error rate in gene data. It is focused to enhance ASD classification model's accuracy. The Enhanced Convolutional Neural Network algorithm is used for ASD classification. The autism microarray dataset from the benchmark public repository, Gene Expression Omnibus (GEO) (National Center for Biotechnology Information (NCBI) is used for analysis. The proposed work using ISSA and ECNN exhibits better performance in terms of precision, accuracy, specificity, sensitivity and time complexity as indicated in the results.

Key Words: Autism Spectrum Disorder (ASD), Gene Selection, Improved Salp Swarm Algorithm (ISSA), Classification, Enhanced Convolution Neural Network (ECNN).

DOI Number: 10.14704/nq.2021.19.9.NQ21142 **NeuroQuantology 2021; 19(9):97-109**

Introduction

A group of neuro developmental disorders is Autism Spectrum Disorder (ASD) which is characterized by communication skills and social interaction difficulties. Repetitive behaviors and interests are limited and various genome-wide association studies (GWAS), meta-analyses, genetic association studies, various genes having association with single nucleotide polymorphisms (SNPs) are identified for ASD. However, from population to population, there will be variations in association among various ASD and SNPs. Changes in diagnostic techniques and screening tools usage as well as changes in various epidemiological techniques has tremendously increased. Estimated ASD incidences have increased awareness among healthcare professionals and general population. In ASD occurrence, unusual increase is caused by various other factors (Sun et al., 2015).

ASD affects one out of every 59 children in the United States, according to the most reliable report (Baio et al., 2018).

Corresponding author: Dr.J. Anitha
Address: [1*]Associate Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore, India.
[1*]E-mail: anitha.j@srec.ac.in
Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
Received: 06 May 2021 **Accepted:** 21 August 2021

Significantly, there have been variations in the frequency of ASD among different ethnic groups. For example, the approximate prevalence of ASD in South Korea is 2.64 percent (Kim et al., 2011), and in China, it is about 12 per 10,000. Since there are differences in the prevalence of ASD among different populations, comparing estimates of ASD prevalence in different regions is challenging. This may be attributed to the various case recognition processes. Males are diagnosed with ASD more often than females, with boys having a 4- to 7-fold greater chance of developing autism than girls; however, the cause for this is still unclear (Werling & Geschwind 2013). Currently, ASD is diagnosed using behavioral criteria that describe deviations from a normal behavior pattern, but what constitutes a typical behavior pattern varies by community. As a result, cultural norms and values will affect the degree of divergence from normal behaviors in ASD (Norbury & Sparks 2013).

In the etiology of ASD, complex genetic factors play a significant role. Genetic diseases are closely linked to ASD, according to family reports. Early research found an 8–10% recurrence risk in siblings with ASD probands, with more recent studies finding up to 25% of siblings affected (Constantino et al., 2010). It means that the likelihood of an ASD child's siblings is about 20 times greater than the risk of the general population. For identifying gene-associated diseases like ASD and genetic disorders, gene expression data having association with diseases are focused by biologists. They have used a high-cost as well as time-consuming techniques like Whole Exome Sequencing (WES), Copy Number Variation Studies (CNV) and Genome-Wide Association Studies (GWAS) (Lee et al., 2012); (Timothy et al., 2013). For candidate disease genes prioritization or prediction, highly reliable as well as low-cost solutions are provided by computational techniques. Associated diseases gene functional information and diverse data sources are combined in computational techniques for making disease genes prediction according to machine learning techniques. Parents experience model with ASD is illustrated in figure 1.

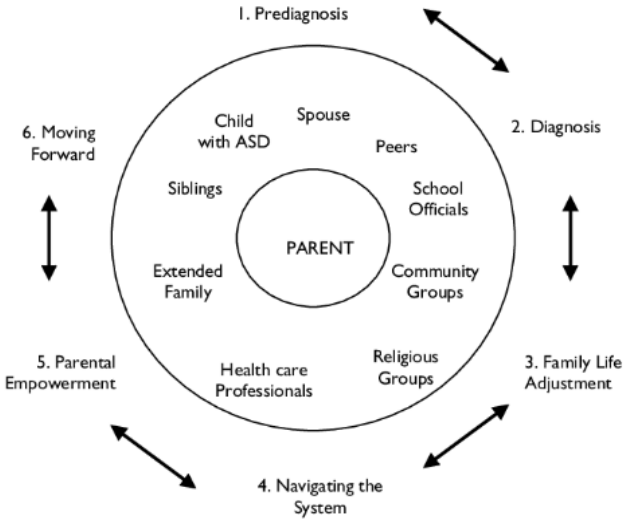

Figure 1. Model's of Parents Experience with ASD

Hidden information is obtained using feature extraction, which are utilized for identifying presence of ASD symptoms in children. A type of pattern recognition problem is autism detection which are having influence of two factors namely, classifier and feature extraction. Class's discrimination process is facilitated using extracted features. Normal classifier like mean distance is enough for this. At early stage, EEG signal analysis is used for disorder detection and identifies ASD as soon as possible (Ibrahim et al., 2018). In other scenarios, for getting better results, classifier and features compatibility are required. From input data which may be a text in specific condition, derived values called features are collected using feature extraction and distinctive properties are generated. Machine learning algorithms generalization and learning ability are enhanced using this informative property. There is a features subset which has highly relevant information.

Highly effective as well as discriminant feature subset are computed using an important stage called feature selection process and datasets quality is enhanced using this and faster as well as better results are produced. For presenting autism risk gene's genome-wide prediction, human brain-specific gene network based complementary machine-learning technique is developed in (Krishnan et al., 2016) which includes more candidates for which there is no prior genetic evidence or minimal. Small key pathways count and brain's developmental stages are converged using large ASD genes set as demonstrated by leveraging brain-specific network and genome-wide predictions. At last, within frequent autism-associated copy-number variants, likely

pathogenic genes are identified and pathways and genes which are most likely to cause ASD across multiple copy-number variants mediators are introduced. For gene selection, an optimization-based algorithm is introduced and from original dataset, significant and relevant features are selected using this.

For a specified input dataset, accurate target class classification is focused by classification methods. For developmental brain gene expression data, with selected features, Support Vector Machine (SVM) model is trained in (Cogill & Wang 2016). For prioritizing candidate autism gene, 74.4% mean sensitivity and 76.7% mean accuracy values are produced using this. An individual brain network is constructed in (Kong et al., 2019) as feature representation and for performing ASD/TC classification, a Deep Neural Network (DNN) classifier is used. At first, for every subject, an individual brain network is constructed by this and between every ROIs pair, connectivity features are extracted. Then using *F*-score, in descending order, these connectivity features are ranked and selected only top ranked features. Finally, 3000 top features are selected for ASD classification and DNN classifier is used.

Autism Spectrum Disorder (ASD)'s risk gene classification is a major problem in this research. Various methodologies and research are introduced for this but their ASD identification accuracy are not up-to the required level. Inaccurate ASD risk classification results and time consumption are the major drawbacks of available techniques. For solving these issues, proposed an Enhanced Convolution Neural Network (ECNN) and Improved Salp Swarm Algorithm (ISSA) is proposed in this research which enhanced overall performance in prediction. ECNN based classification, ISSA based gene selection and pre-processing are deployed in this proposed work to provide improved and better results.

The rest of the paper is organized as follows: a brief review of some of the literature works in preprocessing, gene selection, and classification methods on ASD are presented in Section 2. Proposed methodology for ISSA+ECNN is detailed in Section 3. Experimental results and performance analysis discussion is provided in Section 4. Finally, conclusions are summed up in Section 5.

Related Work

In children having autism spectrum disorder (ASD) and autism (AU), in blood, gene expression differences are identified by Gregg et al., (2008) and comparison is made with general population controls. With gender and age matching of typically developing children from general population, comparison is made between transcriptional profiles.

According to early onset (A–E without regression) history or developmental regression (A–R) history, the AU group is subdivided. All genes are expressed differentially and in autism, emerging evidence for abnormalities in peripheral blood leukocytes is supported by gene expression data as concluded.

From every subgroup related with control group, significant differentially expressed genes are discussed by Hu et al., (2009). In common across subgroups, differentially expressed genes unique to every genes as well as subgroups, it reveals. As concluded, there are 15 genes which are severely affected in ASD group which are responsible for regulating circadian rhythm, which has various effects on metabolic as well as neurological functions.

Among all three ASD subgroups, there are *20 novel genes* which are located in noncoding regions. With androgen sensitivity they are having association and they may have strong 4:1 bias towards affected males.

On Bayesian technique, a new discretization technique is introduced by Boullé (2006). On this model space, a prior distribution and discretization models' space is introduced. In data mining, most important data preprocessing technique is discretization process. With a discretization algorithm, other algorithms can be integrated if attributes are continuous and they are transformed as discrete feature.

In discretization's Bayes optimal evaluation criterion is defined using these results. A new super-linear optimization algorithm is used and near-optimal discretization are computed. New discretization technique produces high inductive performance as demonstrated in the experimentation done on synthetic and real data.

In class imbalance problem, support vector machine (SVM) application is dealt by Farquad & Bose (2012). As a preprocessor, SVM's efficiency and feasibility are examined in this paper. SVM is employed as a preprocessor and trained SVM's predictions are used for replacing training data's

actual target values.

As per sensitivity measure, data is effectively balanced using proposed technique as observed and for minority class, more instances are provided using this technique and Intelligence technique's performance are enhanced using this.

In impulse noise presence, for optical character recognition, a novel algorithm is used by Spitsyn et al., (2016). Neural networks, Principal Component Analysis and Haar wavelet transform are applied in this work. Feature extraction, noise elimination and low frequency components allocation are done using Haar wavelet transform. Extracted features dimension are minimized using principal component analysis.

For every character, various multi-layer neural networks are used in this work. Reduced feature set is used for representing inputs. For every character type, a separate neural network is created in proposed technique, which is key feature of this. With impulse noise, image characters are recognized effectively using this proposed algorithm as indicated in experimental results.

According to chaos theory and SSA, a novel hybrid solution is introduced by Sayed et al., (2018). On 20 benchmark datasets, and on14 unimodal and multimodal benchmark optimization problems, applied this Chaotic Salp Swarm Algorithm (CSSA). For enhancing precision and convergence rate, employed various ten chaotic maps. Proposed CSSA is a promising algorithm as shown in simulation results.

In an optimal feature subset computation. CSSA's capability is revealed in experimentation results. Classification accuracy is enhanced and selected features count is minimized using this algorithm. Optimum map is logistic chaotic map as shown in results which enhances original SSA's performance significantly.

In autistic individual, for ranking available diseased genes set, an approach is used by Reeta et al., (2018). In training set, similarities between diseased and individual genes are compared by the system for predicting individual's autistic behavior. Naïve Bayesian classification technique is used for implementing this.

For example, in individual's DNA test, if DNA has diseased genes which are available in training set, then autism is predicted. At an early stage, autism diagnosis is made easier using this technique.

In ASD diagnosis, DNN model's performance is analyzed by Misman et al., (2019). Using two adult ASD screening datasets, with respect to classification accuracy, this analysis is performed. Previous Machine learning technique called Support Vector Machine (SVM) is used for comparing the results.

On first dataset, in ASD diagnosis classification, around 99.40% accuracy is achieved using DNN model and on second dataset, around 96.08% is achieved. At the same time, 95.24% and 95.08% is achieved by SVM model as shown in experimentation results. The DNN classification technique is implemented using ASD adult screening data for accurately identifying ASD cases as shown in results.

Proposed Methodology

For enhancing ASD risk gene classification, this work proposes an Enhanced Convolutional Neural Network (ECNN) and Improved Salp Swarm Optimization (ISSA) classifier. Gene classification, gene selection and data pre-processing are involved in this proposed work. Figure 2 shows the proposed system's overall block diagram.

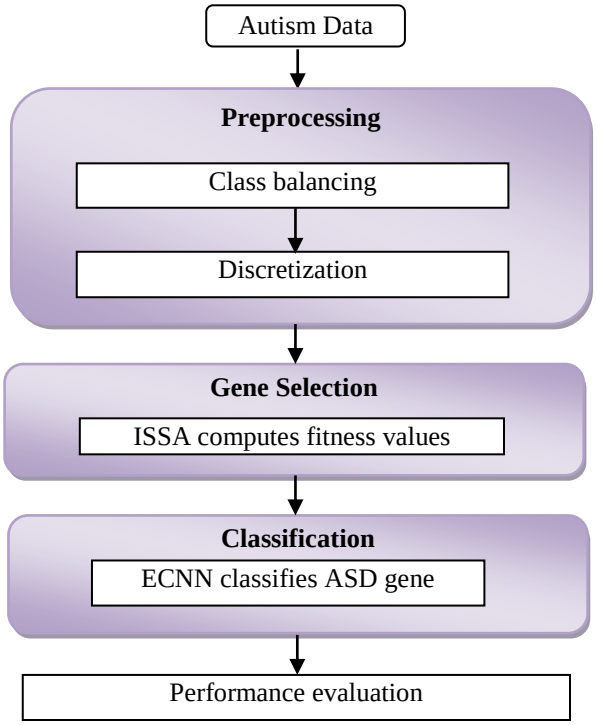

Figure 2. Overall Block Diagram of the Proposed System

Data Pre-Processing

Discretization and class balancing techniques are included in this model's preprocessing stage. In a dataset, features are described, quantified and computed using data pre-processing techniques.

Machine learning algorithm's performance like accuracy and learning time are enhanced using this proposed technique. Three data pre-processing techniques are combined in this work. For ASD risk gene prediction, more robustness is shown by this work.

Class Balancing

In data mining, major challenging research problem is imbalanced data (Yang & Wu 2006). In a two-class dataset, few instance in one class which may be a target class than other class causes imbalanced data. In this scenario, classification algorithm may be misleading by majority class with overwhelmed data.

In can also be stated as, classifier may be over-fitted and in specific, on minority class, its performance is degraded (Sandberg 2000). In general, misclassification of minority class samples occurs very often than majority class. For fixing this problem, re-weighting of instances is done which preserves total weights sum across all instances and every class's total weight.

Discretization

A type of data preprocessing which is used for minimizing values count for a specified continuous attributes is termed as Discretization. In this, attribute's range is split as intervals. More accurateness, shortness and compactness are shown by discrete features than continuous features.

In specific, there are four steps in normal discretization process. They are, all feature's continuous values which needs to be discretized are sorted, cut point selection for splitting continuous values as intervals, continuous value's intervals merging or splitting, discretization process's stopping criteria selection. The lncRNA gene data are optimized after preprocessing stage as illustrated in table 1.

Gene selection using Improved Salp Swarm Optimization (ISSA) with Enhanced Convolutional Neural Network (ECNN) Algorithm for ASD Prediction

For a specified dataset, optimum gene selection is done using ISSA algorithm in this work. A recently introduced metaheuristic algorithm based on swarm is Salp Swarm Algorithm (SSA) (Mirjalili et al., 2017). In oceans, salps swarming behavior is mimicked in SSA. This behavior is used for food source searching and navigation.

A type of stochastic algorithm is SSA where initial random solutions set is created for initiating initial population which is used for starting optimization process. Over the time, these solutions are enhanced in two phases namely, exploitation or intensification and diversification or exploration. For discovering promising regions, search space is explored in exploration and specific solution's neighborhood are searched in exploitation and are used for computing better solution than current solution.

Due to their simplicity, flexibility, performance in computing better solutions to real world problems and ability in avoiding local optimum value, more popularity is gained by nature inspired algorithms (Mafarja et al., 2017). This is because, from real world behaviours like physics, human, swarms, etc., they are inspired. In nature, Salps behavior is mimicked in SSA. Creatures living in oceans and seas are called salps. Tissues as well as movement towards food sources of Jellyfishes and salps are similar. Group of salps is termed as Salp chains. There is a leader and followers set in every chain as illustrated in figure 3.

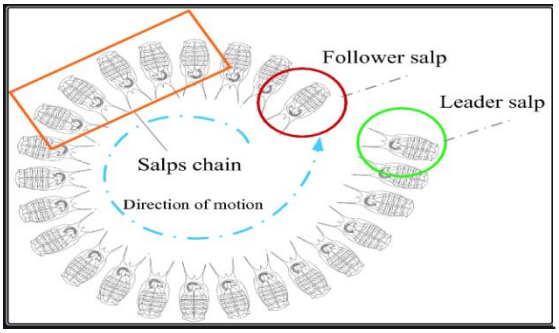

Figure 3. Nature of SSA Algorithm

Salp chain is split into two groups namely followers and leader by swarming behavior of mathematical model. In chain, leader corresponds to first salp and followers corresponds to remaining salps. In salp chain, chain is leaded by leader and followers follow each other. In SSA, salp chain corresponds to population and in chain, salp position is represented using every solution in population.

There are n-dimensions in every solution, where, problem variables count is represented as n. Thus, all salps positions are stored using two-dimensional matrix. In population, best solution is expressed as target or food source and it is represented as F.

In optimization process, in salp chain, based on

every solutions role, they are updated. Salps positions are updated using following expression which includes followers and leaders. Following expression is used for updating leader's position.

$$x_j^1 = \begin{cases} F_j + c_1\left((ub_j - lb_j)c_2 + lb_j\right) & c_3 \geq 0 \\ F_j - c_1\left((ub_j - lb_j)c_2 + lb_j\right) & c_3 < 0 \end{cases} \quad (1)$$

Where, lead j^{th} dimension's position is represented as x_j^1 j.

Best food source or solution is represented as F_j, in jth dimension, lower bound is given by lb_j and upper bound is given by ub_j. Random variables are given by c1, c_2 and c_3. In SSA performance, vital role is played by c_1. Between exploitation and exploration, balance is controlled using this parameter.

As indicated in expression (1), iteration count based time varying parameter is represented as c_1. At early optimization process stages, high exploitation rates are allowed using this. In last stages, high exploitation rates are allowed.

$$c_1 = 2e^{-(\frac{4l}{L})^2} \quad (2)$$

Where, current iteration is represented as l, maximum iterations count is represented as L, uniform random numbers are represented as c_2 and c_3, their value will lie between 0 to 1. In addition to step size, movement of next position in j th dimension towards $+\infty\ or\ -\infty$ are indicated using these variables.

Follower's positions are updated by simulating Newton's law of motion which is expressed in (2).

$$x_j^i = \frac{1}{2}(x_j^i + x_j^{i-1}) \quad (3)$$

Where $i \geq 2$ and eat jth dimension, ith follower's position is depicted as x_j^i.

In SSA, towards food source, leader salp will move and towards leader, follower will move. During the process, there is a possibility to change food source position and leader will move towards new food source position. However, local optima trapping and population diversity problems are there in SSA as like other optimization procedures.

For resolving those problems, SSA's enhanced version called Improved Salp swarm algorithm (ISSA) is presented in this research. For solving its problems, two major enhancements are included in this SSA. For salps' position update, new expression is developed, which is a first enhancement. This improvement enhances SSA solutions' diversity. A new Local Search Algorithm (LSA) is developed as a second enhancement which is used for enhancing SSA exploitation ability.

Algorithm 1: LSA Algorithm

$T_1 = F$
$K = 1$
While (K<max_LSA_number of itertaions)
Selected features count = rand of 2 or 5
From T_1, 2 or 5 features are randomly selected by LSA as per selected features values count
For every selected features in T_1
If feature value = = 1 (1 indicates feature selection and o means ignorance)
Feature value =0
Else
Feature value = 1
End if
T_1's fitness value is computed
If $f(T_1) < f(F)$
$F = T_1$
End if
$K = K + 1$
End while
Return F

Solution's diversity is enhanced using first enhancement and in SSA, between exploration and exploitation, balance is made using this. LSA algorithm's conditional usage is included in second enhancement, which is used for enhancing current best solution and SSA's exploitation ability. At every SSA iteration's end, LSA is called on current best solution F, for checking the availability of better solution. However, for avoiding more computational requirements, at every iteration, LSA is not used.

Therefore, for tracking best solution F value's enhancement, at SSA end, a new count variable called "Improvement_counter" is included. This value indicates SSA iterations count which occurred without F value's enhancement. For example, if "Improvement_counter ≥ 2'', it indicates, two iteration are not having any F value enhancement. So, for enhancing current best solution F's value, LSA algorithm is called. Necessary initializations are made at LSA beginning.

At every LSA iteration, according to "Number of selected features" variable value which are randomly generated, selected two or five features. The "Number of selected features" variable value will be either two or five for covering large as well as small datasets.

Now, selected features value is flipped to either 1 (feature selection) or 0 (feature ignorance). In addition, new solution T_1's fitness value is

computed using LSA and if current F best solution is worst than this, F value is updated by LSA and its value is assigned to T_1 value.

Algorithm 2 shows ISSA's overall enhancement which displays proposed ISSA algorithm's pseudo code according to mentioned enhancements.

Algorithm 2: ISSA Algorithm

1. Salps (searching genes) positions x_i $(i = 1,2,3, \dots n)$ are initialized
2. In chain, every gene or salp's fitness gene are computed
3. F =best salp position according to fitness value
4. $While(t < max + iterations)$
5. Using expression (2), c_1 is updated
6. For (every salp(x_i))
7. $if\ (i == 1)$
8. Using expression (1), leader gene or salp's positions are updated
9. Else
10. Using expression (3), follower salp positions are updated
11. End if
12. End for
13. Genes which go further than search space splits are repositioned
14. In chain, every autism's fitness values are computed
15. F= best gene position according to fitness value
16. Improvement counter variable's value is computed
17. If (improvement_ counter≥2)
18. Algorithm 1 (LSA) is called on F for enhancing its value
19. End if
20. $t = t + 1$
21. End while

*Return **F** (Best Genes)*

Two enhancements are included in this ISSA algorithm. For followers' position update, expression (4) is used for selecting best local genes. LSA usage is another enhancement, where, according to "Improvement_counter" variable values, LSA is executed conditionally. In initial ISSA stage, random slaps or population count is generated. Using a binary value, every solution called Salp are represented. Selected features are represented as "1'' and not selected features are represented as "0''.

Also, a multi-objective optimization problem can be formulated from gene selection problem. In solution, with minimal selected or included genes count, high classification accuracy is concentrated in this. Then, according to fitness function, every solution's food fitness is computed in ISSA. Now, current best solution or food source corresponds to the solution having minimum fitness value.

Therefore, best solutions so called food sources are selected by ISSA using guidelines of fitness objective function. In next stage, slaps positions are updated by ISSA using expression (2) and (4). In every iteration, for updating best solution or food source, fitness values are recomputed. At last, if maximum iterations count is attained by ISSA, best gene solution is returned as binary values vector.

For a specified dataset, for providing highly accurate ASD risk gene classification results, ECNN is proposed in this work. For classifying test data as no or yes classes, this proposed work introduces Enhanced Convolutional Neural Network (ECNN). High accuracy is attained using proposed deep learning technique. There are multiple hidden layers, output and input layers in basic CNN.

In general, CNN's hidden layers have fully connected, pooling and convolutional layers. Input is applied with convolution operation in convolutional layers and to next layer, results are transferred. Individual neuron's response is emulated as visual stimuli using convolution operation.

Global or local pooling layers are included in convolutional networks. In one layer, neuron cluster's outputs are combined as a single neuron in next layer using this. In previous layer, every neuron cluster's average value is used by mean pooling. In one layer, every neuron is connected to every neuron in next layer using fully connected layers.

Principle of traditional multi-layer perceptron neural network is similar to CNN (Suganuma et al., 2017); (Li et al., 2018). Classification, sub-sampling, convolutional and input layers are there in proposed ECNN. In high-dimensional data analysis, this proposed technique shows its effectiveness. Parameter sharing scheme is employed by this. Parameters count are minimized as well as controlled using these convolutional layers.

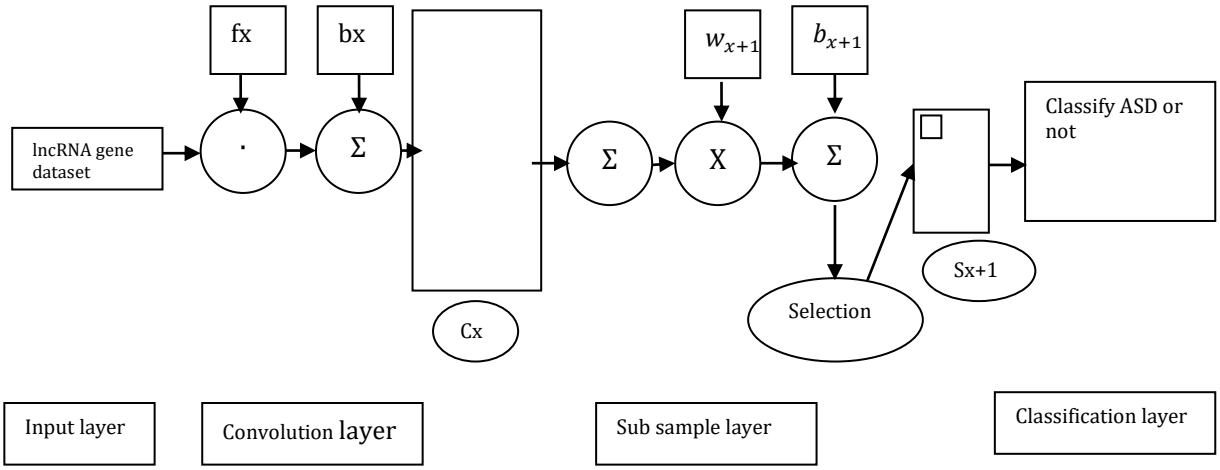

Figure 4. Architecture Diagram of ECNN

From training samples, query images called features are given to input layer and data is transformed as a unified form for delivering it to next layer properly. Initial parameters like various filters, local receptive field's scale value are defined in this layer. Figure 4 illustrates the ECNN's architecture diagram.

Via convolution algorithm, input data are processed by Convolution layer (Cx) and various layers termed as feature map are produced in this. These maps have convolution computation results of previous layers. Key features extracted and network's computational complexities are minimized using this.

After every convolutional layer, employed an activation function. Outputs are mapped to input set using activation function and network structure is made more non-linear because of this. For entire specified feature values, initial connection weights are assigned. Then, applied a new input pattern and output is calculated as,

$$y(n) = f(\sum_{i=1}^{i=N} w_i(n)x_i(n) \qquad (4)$$

$$\text{Where } f(x) = \begin{cases} +1 \; if \; x \geq 0 \\ -1 \; if \; x < 0 \end{cases} \qquad (5)$$

Where, iteration index is represented as n.

Based on following expression, updated connection weights,

$$w_i(n+1) = w_i(n) + \eta(d(n) - y(n)x_i(n), i = 1,2,..N \qquad (6)$$

Where, gain factor is represented as η

Then standard deviation is applied

$$\sigma = \sqrt{\frac{1}{n} f_i(x_i - \overline{x)^2}} \qquad (7)$$

Proposed ECNN network is given with these weighted gene dataset features and highly accurate classification results are obtained. On same dataset, major analysis findings are confirmed using polynomial distribution function. In this layer, sub-sampling of every feature map from previous convolution layer is done.

Informative features sum is given by $Sx + 1$ as shown in Figure 4. This DLCNN algorithm is combined with genetic algorithm for selecting gene values with high accuracy, which enhances overall ASD detection performance.

Last layer is highly sensitive to connection weight with previous layer as it is responsible to compute classification result and it determines current image's belonging level for ten classes from "0" to "9". Last weight vector's training is optimized in this work via effective ISSA algorithm, which enhances classification accuracy.

In swarm, global best and local best positions are computed using ISSA. In training process, using expression (3), genes or slaps are updated. In dataset, based on final global best genes, updated output weight vector at ISSA iterations end. Then, for computing classification rate, test dataset is used for testing trained model.

Algorithm 3: Steps in ECNN

Input: testing data, training data, population size, lncRNA gene dataset

Objective function: Prediction accuracy is assigned as fitness function

Output: ASD or Non ASD genes

1. Start
2. Procedure lncRNA gene dataset
3. For every input attribute, gene attribute ∈ lncRNA gene dataset is described do
4. Population is initialized
5. While t<max generation count do
6. Input is converted into sub layers

7. Using (6) and (7), non ASD and ASD genes are detected
8. Call Algorithm 2
9. Using ISSA algorithm, highly relevant as well as informative features are selected.
10. For specified dataset, testing and training process are performed
11. $t = t + 1$
12. Highest detection accuracy features are returned
13. For every feature, based on input dataset, predefined gene labels are copied.
14. End

Numerous genes indications are learned for classifying ASD risk gene detection results more accurately.

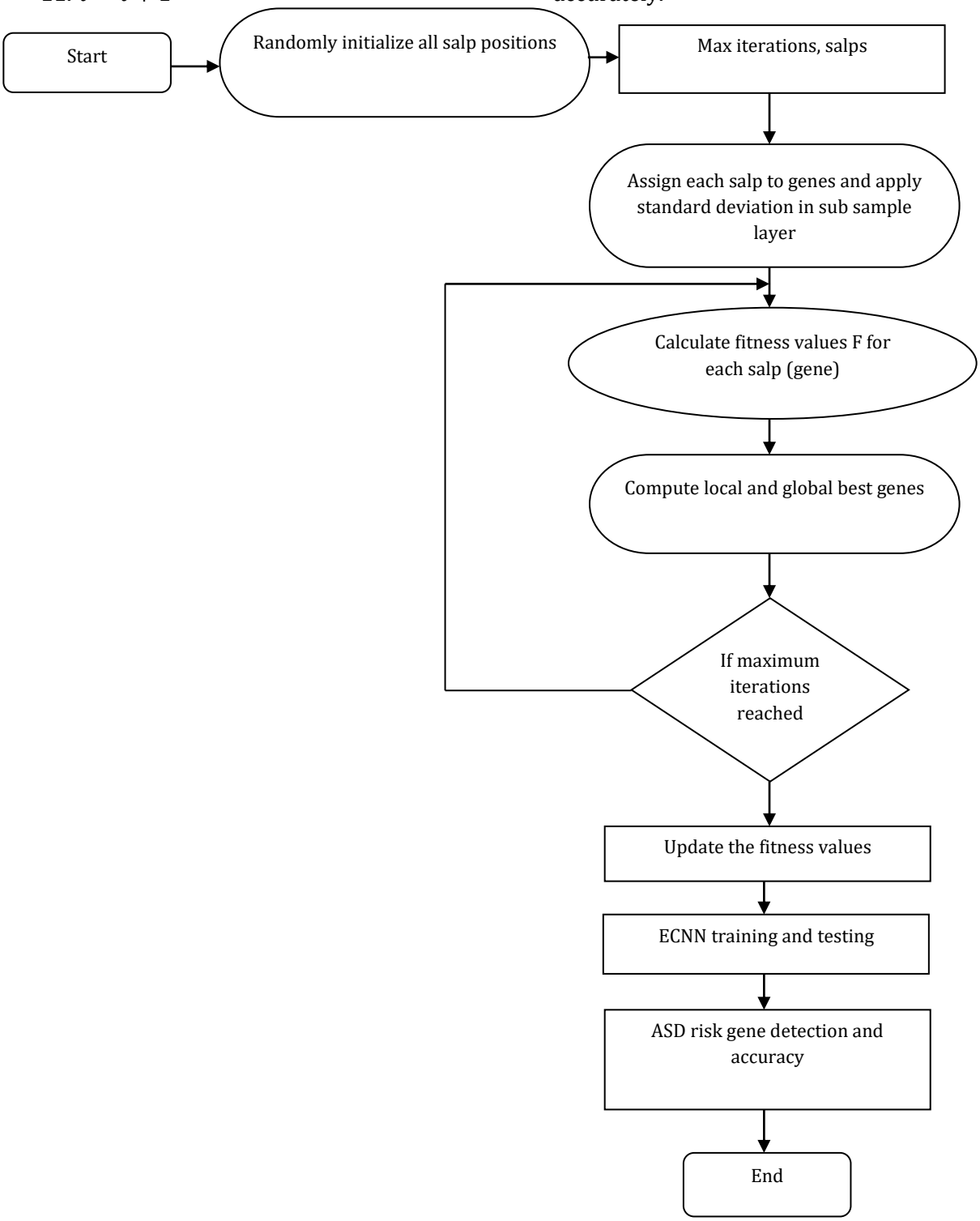

Figure 5. Overall Flow Diagram of the Proposed ASD Gene Detection Process

For classifying specified lncRNA gene dataset, best ECNN architecture is discovered in this proposed system. On a specified lncRNA gene dataset, every individual who encodes a specific ECNN's architecture's fitness is evaluated in evolution process. According to this fitness, best genes are selected and in testing and training process, ECNN classification accuracy is enhanced using this.

Experimental Result

This experimentation employs a data which are utilized in (Cogill & Wang 2016). From The BrainSpan Atlas of the Developing Human Brain developmental transcriptome dataset, extracted this data with IncRNA as instances and for tem-porospatial time points, features corresponds to expression values. There are 366 known ASD genes called positive data, which are spanned from 2128 disease genes with 26 brain structure's 524 respective developmental time points. The age limit lies between 8 weeks to 40 years.

Precision

Precision value is computed as,

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}} \quad (8)$$

Quality or accuracy is computed using precision value and quantity or completeness is measured using recall value. Computation of highly relevant results than irrelevant results by an algorithm is indicated using high precision value. For a class, ratio between true positives count to total elements which are labelled as positive class's count defines precision value in classification task.

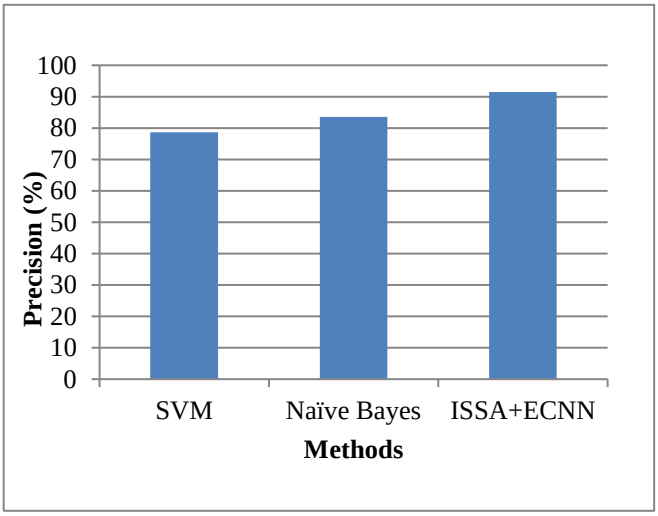

Figure 6. Precision

Precision performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 6. In x axis of the above mentioned plot, various techniques are considered and precision values are considered in y axis. For a specified autism dataset, high precision value is provided by proposed ISSA+ECNN algorithm, whereas, low precision value is provided by available techniques like Naïve Bayes and SVM. In lncRNA gene dataset through optimal genes selection, ASD classification accuracy is enhanced using proposed ISSA+ECNN as concluded in results.

Sensitivity

It is also termed as true positive rate, detection probability or recall. Actual positives proportion which are identified correctly as such defines this value. It can also be stated as sick people's percentage who are identified correctly as with condition.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (9)$$

Where, True Positive is expressed as TP and False Negative is expressed as FN.

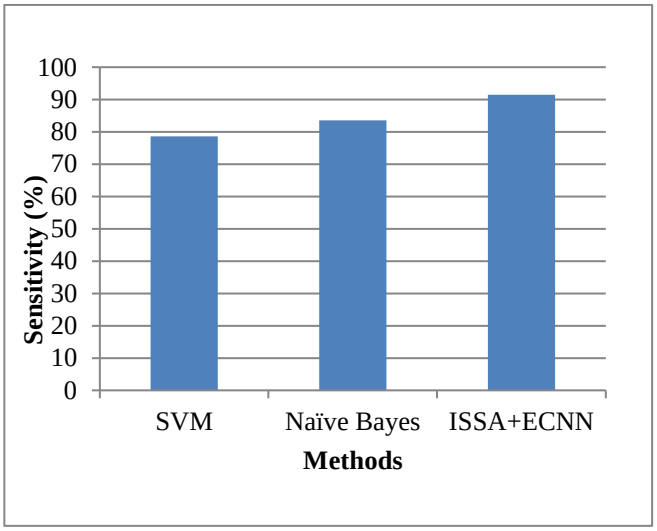

Figure 7. Sensitivity

Sensitivity performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 7. In x axis of the above mentioned plot, various techniques are considered and sensitivity values are considered in y axis. For a specified autism dataset, high sensitivity value is provided by proposed ISSA+ECNN algorithm, whereas, low sensitivity value is provided by available techniques like Naïve Bayes and SVM. In lncRNA gene dataset, through optimal genes selection, ASD classification accuracy is enhanced using proposed ISSA+ECNN as concluded in results.

Specificity

Specificity is also termed as true negative rate. Actual negative's proportion which are identified correctly as such defines this specificity value. It can also be stated as healthy people's percentage who are identified correctly as not with condition.

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (10)$$

Where, True Negative is expressed as TN and False Positive is expressed as FP.

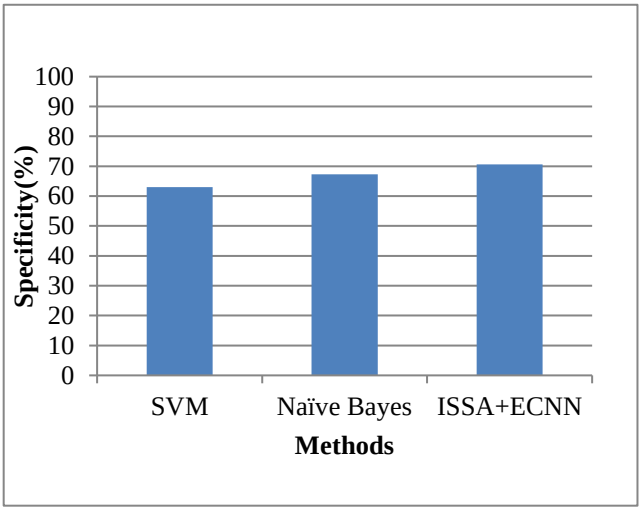

Figure 8. Specificity

Specificity performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 8.

In x axis of the above mentioned plot, various techniques are considered and specificity values are considered in y axis. For a specified autism dataset, high specificity value is provided by proposed ISSA+ECNN algorithm, whereas, low specificity value is provided by available techniques like Naïve Bayes and SVM.

In lncRNA gene dataset, through optimal genes selection, ASD classification accuracy is enhanced using proposed ISSA+ECNN as concluded in results.

Accuracy

Model's overall correctness is computed as accuracy and ratio between total actual classification parameters $(T_p + T_n)$ and classification parameters sum $(T_p + T_n + F_p + F_n)$ defines this accuracy value. It is expressed as,

$$\text{Accuracy} = \frac{T_p+T_n}{(T_p+T_n+F_p+F_n)} \qquad (11)$$

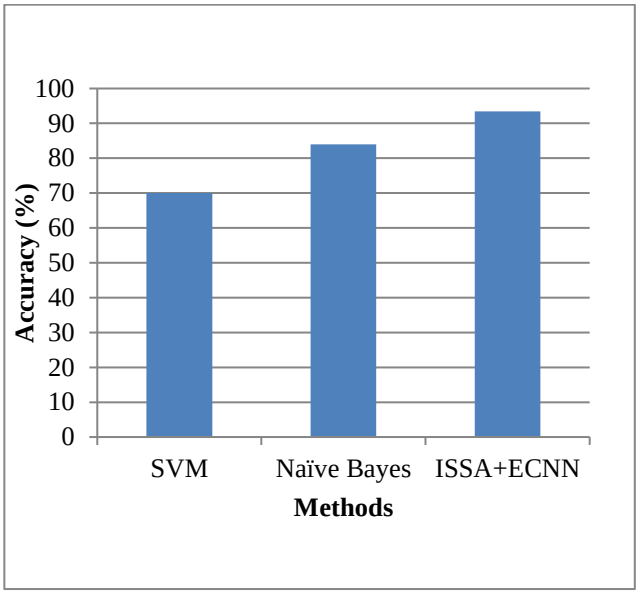

Figure 9. Accuracy

Accuracy performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 9. In x axis of the above mentioned plot, various techniques are considered and accuracy values are considered in y axis.

For a specified autism dataset, high accuracy value is provided by proposed ISSA+ECNN algorithm, whereas, low accuracy value is provided by available techniques like Naïve Bayes and SVM. Through optimal genes selection, ASD classification accuracy is enhanced using proposed ISSA+ECNN as concluded in results.

False Positive Rate (FPR)

Ratio between negative events count which are categorized wrongly as positive and total actual negative events count irrespective of classification defines false positive rate. False positive ratio's expectancy is referred as false positive rate or false alarm rate.

$$\text{The false positive rate} = \frac{FP}{FP+TN} \qquad (12)$$

Where, false positives count is represented as FP, true negatives count is represented as TN and total ground truth negatives count is given by N=FP+TN.

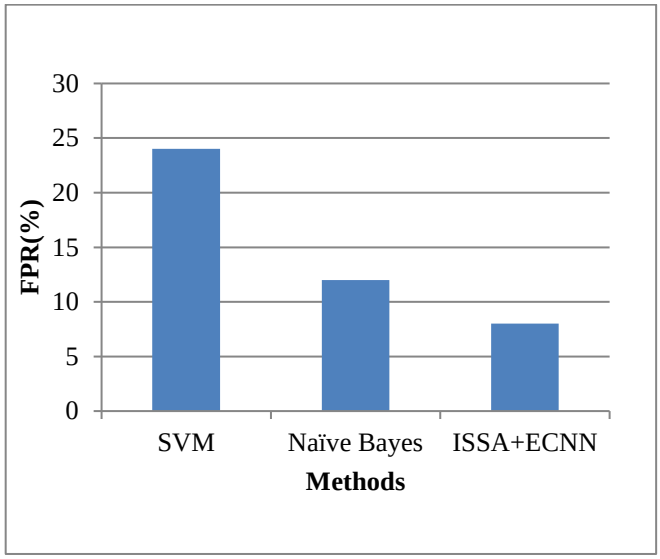

Figure 10. FPR

FPR performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 10. In x axis of the above mentioned plot, various techniques are considered and FPR values are considered in y axis. For a specified autism dataset, low FPR value is provided by proposed ISSA+ECNN algorithm, whereas, high FPR value is provided by available techniques like Naïve Bayes and SVM. Through optimal genes selection, ASD classification accuracy is enhanced using proposed ISSA+ECNN as concluded in results.

Time Complexity

If less time is consumed for executing proposed algorithm, then the system exhibits better response.

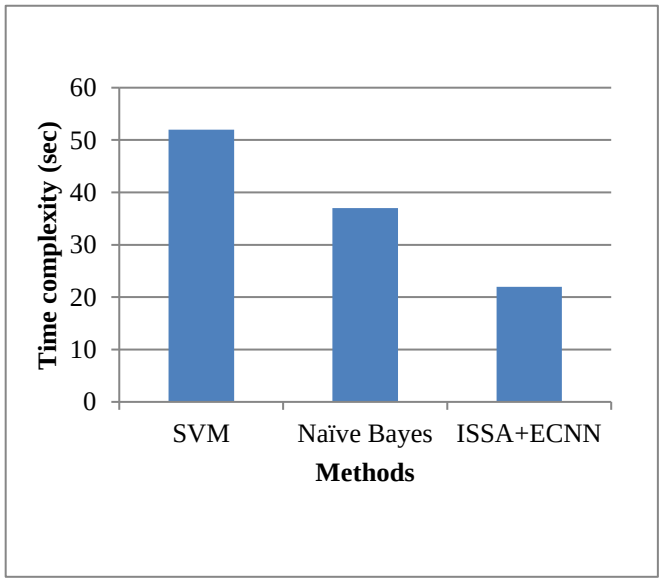

Figure 11. Time Complexity

Time complexity performance evaluation metric comparison between proposed and available technique is illustrated in above mention figure 11. In x axis of the above mentioned plot, various techniques are considered and time complexity values are considered in y axis. For a specified autism dataset, low time complexity value is provided by proposed ISSA+ECNN algorithm, whereas, high time complexity value is provided by available techniques like Naïve Bayes and SVM. In ASD, results are accurately classified using proposed ISSA+ECNN which enhances classification performance as concluded in results.

The above mentioned performance metrics comparison between proposed and available technique in lncRNA gene dataset is shown in table 1.

Table 1. Comparison Values for in cRNA Gene Dataset

Methods/Metrics	SVM	Naïve Bayes	ISSA+ECNN
Accuracy (%)	70	84	93.4
Precision (%)	78.6	83.5	91.45
Sensitivity (%)	65	68	72
Specificity (%)	63	67.3	70.67
FPR (%)	24	12	8
Time complexity (sec)	52	37	22

Conclusion

In recent years, it is essential to diagnosis ASD accurately as it is required for its rehabilitation and management. A type of nervous system disorder is ASD which affects brain and leads to difficulties in motor abilities delay, repetitive behaviors, communication and social interaction deficits and speech difficulties. For enhancing ASD risk gene classification accuracy, ISSA+ECNN algorithm is proposed in this research. From specified autism dataset, irrelevant features are removed in preprocessing stage. Class balancing and discretization is used for the same. Then, through ISSA algorithm, gene selection is done. Relevant autism features and robust genes are computed using fitness features. Then ECNN algorithm is used for performing classification. Through genetic algorithm's optimal fitness values, ASD detection accuracy is enhanced because of this classification algorithm. High precision, accuracy, specificity, sensitivity and low FPR, time complexity results are produced by proposed ISSA+ECNN algorithm as demonstrated in experimental results than other available techniques. In future work, Fuzzy based soft computing techniques can be developed for improving the FPR over larger number of datasets.

The scope of autism services research needs to broaden to include the full array of services for adults across domains, including housing, health care, criminal justice, and financial planning applications.

References

Sun X, Allison C, Auyeung B, Zhang Z, Matthews FE, Baron-Cohen S, Brayne C. Validation of existing diagnosis of autism in mainland China using standardised diagnostic instruments. *Autism* 2015; 19(8): 1010-1017.

Baio J, Wiggins L, Christensen DL, Maenner MJ, Daniels J, Warren Z, Dowling NF. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries* 2018; 67(6): 1-23.

Kim YS, Leventhal BL, Koh YJ, Fombonne E, Laska E, Lim EC, Grinker RR. Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry* 2011; 168(9): 904-912.

Werling DM, Geschwind DH. Understanding sex bias in autism spectrum disorder. *Proceedings of the National Academy of Sciences* 2013; 110(13): 4868-4869.

Norbury CF, Sparks A. Difference or disorder? Cultural issues in understanding neurodevelopmental disorders. *Developmental psychology* 2013; 49(1): 45-48.

Constantino JN, Zhang YI, Frazier T, Abbacchi AM, Law P. Sibling recurrence and the genetic epidemiology of autism. *American Journal of Psychiatry* 2010; 167(11): 1349-1356.

Lee KW, San Woon P, Teo YY, Sim K. Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt?. *Neuroscience & Biobehavioral Reviews* 2012; 36(1): 556-571.

Timothy WY, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, Walsh CA. Using whole-exome sequencing to identify inherited causes of autism. *Neuron* 2013; 77(2): 259-273.

Ibrahim S, Djemal R, Alsuwailem A. Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis. *Biocybernetics and Biomedical Engineering* 2018; 38(1): 16-26.

Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, Troyanskaya OG. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature neuroscience* 2016; 19(11): 1454-1462.

Cogill S, Wang L. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics* 2016; 32(23): 3611-3618.

Kong Y, Gao J, Xu Y, Pan Y, Wang J, Liu J. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* 2019; 324: 63-68.

Gregg JP, Lit L, Baron CA, Hertz-Picciotto I, Walker W, Davis RA, Sharp FR. Gene expression changes in children with autism. *Genomics* 2008; 91(1): 22-29.

Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME, Lee NH. Gene expression profiling differentiates autism case–controls and phenotypic variants of autism spectrum disorders: Evidence for circadian rhythm dysfunction in severe autism. *Autism research* 2009; 2(2): 78-97.

Boullé M. MODL: a Bayes optimal discretization method for continuous attributes. *Machine learning* 2006; 65(1): 131-165.

Farquad MAH, Bose I. Preprocessing unbalanced data using support vector machine. *Decision Support Systems* 2012; 53(1): 226-233.

Spitsyn VG, Bolotova YA, Phan NH, Bui TTT. Using a Haar wavelet transform, principal component analysis and neural networks for OCR in the presence of impulse noise. *Computer Optics* 2016; 40(2): 249-257.

Sayed GI, Khoriba G, Haggag MH. A novel chaotic salp swarm algorithm for global optimization and feature selection. *Applied Intelligence* 2018; 48(10): 3462-3481.

Reeta R, Pavithra G, Priyanka V, Raghul JS. Predicting Autism Using Naive Bayesian Classification Approach. *IEEE International Conference on Communication and Signal Processing (ICCSP)* 2018: 0109-0113.

Misman MF, Samah AA, Ezudin FA, Majid HA, Shah ZA, Hashim H, Harun MF. Classification of adults with autism spectrum disorder using deep neural network. *In 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)* 2019: 29-34.

Yang Q, Wu X. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making* 2006; 5(4): 597-604.

Sandberg K. *The Haar wavelet transform.* University of Colorado at Boulder, Boulder 2000.

Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software* 2017; 114: 163-191.

Mafarja M, Eleyan D, Abdullah S, Mirjalili S. S-shaped vs. V-shaped transfer functions for ant lion optimization algorithm in feature selection problem. *Proceedings of the international conference on future networks and distributed systems* 2017: 1-7.

Suganuma M, Shirakawa S, Nagao T. A genetic programming approach to designing convolutional neural network architectures. In Proceedings of the genetic and evolutionary computation conference 2017: 497-504.

Li H, Parikh NA, He L. A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Frontiers in neuroscience* 2018: 1-32.