International Journal of Intelligent Computing and Information Science

ASSOCIATION RULES BASED CLASSIFICATION FOR AUTISM SPECTRUM DISORDER DETECTION

Aya Rashad

College of Computing and Information Technology,

Arab Academy for Science, Technology, and Maritime Transport,

Cairo, Egypt,

ayarashad88@hotmail.com

Fahima Maghraby

College of Computing and Information Technology,

Arab Academy for Science, Technology, and Maritime Transport,

Cairo, Egypt,

fahima@aast.edu

Mohamed Mostafa Fouad

College of Computing and Information Technology,

Arab Academy for Science, Technology, and Maritime Transport,

Cairo, Egypt,

mohamed_mostafa@aast.edu

Yasmin Lashin

Department of Biochemistry

German University in Cairo

Cairo, Egypt,

yasmine.lashine@guc.edu.eg

Amr Badr

Faculty of Computers

and Information,

Cairo University,

Cairo, Egypt

a.badr.fci@gmail.com

Abstract- *Autism Spectrum Disorder (ASD) is a very complicated disorder; a recent study made in the USA showed that it is the second most widespread neurodevelopmental disorder among children. It is a highly undetectable disease since its symptoms are not the same for each child. The traditional detecting methods are not efficient especially in the early years of child development, resulting in late treatment. This paper proposes a machine learning approach that detects ASD through gene signature, at childbirth. The approach processes the data of blood-based gene expression of 21 ASD child and 63 Controls into a feature selection phase using Bagging Tree Algorithm, the delivered features are then weighted through the application of a genetic mathematical formula. A Discretization process is used before the main Association Rules process. The final association rules show the most important relationships between genes for the early prediction of the ASD. The rules showed relationships between 600 genes in which 8 genes are the most affecting ASD with an accuracy of 90%.*

Keywords: Association Rules, Feature Selection, Gene Expression, Bagging Trees, ASD.

I. Introduction

ASD is a group of neurodevelopmental issues described by disabilities in social correspondence and limited interest is a disability that can cause huge social and behavioral difficulties, there is nothing extraordinary in the way ASD patients resemble that separates them from other individuals, however individuals with ASD may very well impart and communicate in ways that are not quite the same as other individuals. The learning, thinking, and problem-solving capacities of individuals with ASD can go from skilled to seriously challenged [1]. Individuals with ASD may repeat certain practices and might not want change in their day by day exercises, numerous individuals with ASD additionally have diverse methods for learning, paying attention, or reacting to things.

The signs of ASD begin during early childhood and last throughout a person's life [1], It can sometimes be detected at 18 months or younger. By age 2, a diagnosis by an experienced professional is considered very reliable, but many children do not receive a final diagnosis until much older and this delay means that children with ASD might not get the early help they need[1].

There is currently no cure for ASD. However, research shows that early intervention treatment services can improve a child's development, it can help children from birth to Three years old (36 months) learn important skills like talk, walk, and interact with others[2].

The Centers of Disease Control and Prevention (CDC)estimate autism's prevalence as 1 in 88 children in the United States. This includes 1 in 42 boys and 1 in 189 girls, and that the ratio of ASD children increases in time [2]. A recent study showed that toddlers, who began behavioral treatment before the age of 3 years, even by the age of 18 months, gained 15 points on a standardized IQ test after the treatment [3]. One of the Methods of detecting it is using the **Gene Expression Analysis,** as most scientists agree that genes are one of the risk factors that can make a person more likely to develop ASD, as Autism Spectrum Disorder is known to have a strong genetic component[8].In order to measure this strong genetic component, Medical tests researches were made to measure it on twins; early twin studies suggested that heritability of ASD was 80-90% [4].

In this paper machine learning technique is proposed to detect ASD, First, the feature Selection is applied to our dataset to select a subset of only the best features in a large number of gene expressions. The bagging tree algorithm was used for such process. After, the dataset was converted into a binary representation in order to apply the association rules technique and produce a list of Biomarker Genes

This paper is organized as follows: Section 1, which is the Related Work and Methods of Diagnosing Autism Spectrum Disorder and section 2 is the detailed explanation of the Research Methodology and used algorithms including Association Rules, Apriori Algorithm and bagging Trees then section 3 explains the Experiment & proposed model & phases section 4 discusses the Result and section 5 discusses the Previous Researches Results and their discussion, finally section 6is a conclusion about the proposed approach

1.Related Work:

Previous researches suggested a complementary machine learning approach using Support Vector Machine (SVM) based on a human brain-specific gene network [5], it constructed a gene-interaction network model containing predicted functional relationships for all pairs within 25,825 genes in the human genome in the brain. The brain-specific network uses a Bayesian method that extracts and integrates brain-specific functional signals from thousands of gene expression,then developed an evidence-weighted, network-based machine learning method that uses this brain-specific network to systematically discover new candidate ASD risk genes across the genome, then trained an evidence-weighted support vector machine classifier using the connectivity of these genes in the human brain-specific network, the Evaluation only held-out through five-fold cross-validation, the approach was accurate (area under the receiver–operator curve (AUC) = 80%

The problem with this study is the limited sample size, as it is unlikely that all risk genes would be identified by sequencing studies and statistical association alone. Sequencing screens mainly identify mutations with large effect sizes, and quantitative association studies rely on relatively high mutation frequencies.

Another research identified the peripheral blood samples of young adults with ASD that can be used to identify a biological signature,where Nineteen differentially expressed probes were identified from a training dataset (Total is 26, 13 ASD cases and 13 controls) using the LIMMA package in R language and were further analyzed in a test dataset (Total=16, 8 ASD cases and 8 controls) using machine learning algorithms,With an overall class prediction accuracy of 93.8% as well as a sensitivity and specificity of 100% and 87.5% [6], the drawback with this research is the application on a small sample size and lack of phenotypic information of the original data. In particular, most ASD subjects in this study exhibited normal intelligence quotients (IQ; mean full-scale IQ, 91.9), this probably does not represent the broader ASD population.

Another study discussed using face scanning methods in order to discover ASD, used a data-driven approach to extract features from the face scanning data and SVM to perform the classification, the evaluation had focused on the performance of the proposed model in terms of its accuracy, sensitivity, and specificity of classifying ASD. Although the results indicated promising evidence for applying the machine learning algorithm based on the face scanning patterns to identify children with ASD, with a maximum classification accuracy of 88.51%, a specificity of 86.21%; and a sensitivity reached93.10%.[7]. However, the face scanning patterns could be age and culture adapted, since people at different ages and in different cultures may scan face differently and also the sample size in the current study is relatively small for the purpose of pattern Classification.

MRI scan is used to identify a small number of FCs (Functional Connectors) that separates ASD child from typically Developed child (TD). Many studies discussed using Brain MRI in ASD detection, one paper [8] did this by establishing a reliable neuroimaging based classifier for ASD by investigating the whole-brain patterns, Using a Machine learning technique applied to the whole set of correlation matrices to optimally select a subset of FCs (Functional Connectors) so that the best classification performance would be obtained. Then applied the $L1$-norm regularized sparse canonical correlation analysis ($L1$-SCCA) to the data set to identify a subset of FCs relevant only to the neural substrates of ASD while factoring out the effects of noise. Then employed the sparse logistic regression (SLR) to further perform dimension reduction to mitigate the over-fitting and thereby extracting the essential FCs representing the core abnormal connectivity in ASD

The classifier separated ASD- from TD (Typically Developed)-populations with an accuracy of 85%. The corresponding area under the curve (AUC) was 93%, indicating high discriminatory ability, but there were some conditions for this classifier to work, the subjects had to be older than 18 years of age, be right-handed, have a IQ exceeding 80, have no comorbidity and have been diagnosed as autistic by either ADI-R or ADOS

The best results were the ones produced from brain [8]and eye tracking studies [9], these were the ones that their finding was most accurate and had a lot of contributions to the ASD community, their only drawback was the fact that they are only applied after the child is 3 years old, the gene expression signature is a much harder and it is still a huge field of experiments to be applied in the future with different techniques and even better findings for it is known to have a strong genetic factor, and to be used since infancy.

1.2 Methods of Diagnosing Autism Spectrum Disorder

There are ways to detect ASD, some are traditional and some are Modern, here are the methods of detecting:

1.2.1 CARS Scale:

- CARS (Childhood Autism Rating Scale) is a 15-item rating scale that evaluates body movements, adaptation to change, listening response, verbal communication, and relationship to people. [10]
- A qualified examiner uses direct observation of a child to identify an autism diagnosis and determine symptom severity.
- The CARS assessment typically takes 5 to 10 minutes.

1.2.2 Eye Tracking:

1. Eye-tracking studies with toddlers with ASD in the age of 3 years have highlighted a range of social visual attention deficits such as a reduced preference for biological motion, reduced fixation to eye and head regions, difficulties in joint attention.
2. GeoPref(Geometric preference)Test: Toddlers are presented with a movie consisting of two rectangular areas that contained Digital geometric images and Digital Social images that were placed side-by-side and scenes changed in a simultaneous, time-linked way identical to our previous experiment, it showed that a subset of toddlers with ASD fixated on geometric images greater than 69%[9]:
3. Toddlers were seated 60 cm in front of the eye tracking monitor. Using a “live tracker” included in the Tobii software (Tobii Studio version 1.3)Tobii is an Eye tracking sensor technology that enables a device to know exactly where your eyes are focused. It determines your presence, attention, focus, drowsiness, consciousness or other mental states, that superimposes the toddler's eye-tracking data on the test image in real time, the operator observed the infant's gaze position and head position on a secondary monitor during the experimental procedure, making note of deviation from an established working range of positions.

Using Tobii software, fixation data were calculated using a 35-pixel radius filter. Time spent fixating and the number of saccades within each area of interest were tabulated for each subject, a receiver operating characteristic curve was generated that graphically displayed a plot of the true positives versus false positives using SPSS statistical software.

1.2.3 Brain MRI:

A Magnetic Resonance Imaging (MRI) scan is a common procedure used by hospitals around the world. MRI uses a strong magnetic field and radio waves to create detailed images of the organs and tissues within the body[9]. As discussed previously in the related work section and in the paper [8], this paper established a reliable neuroimaging based classifier for ASD by investigating the whole-brain patterns and applied multiple Machine learning techniques to extract the essential Functional Connectors representing the core abnormal connectivity in ASD.

1.2.4 Gene Expression signature

Genes are subunits of DNA, the information database of a cell that is contained inside the cell nucleus. This DNA carries the genetic blueprint that is used to make all the proteins the cell needs. Every gene contains a particular set of instructions that code for a specific protein, Gene expression is the process by which genetic instructions are used to synthesize gene products. These products are usually proteins, which go on to perform essential functions as enzymes, hormones, and receptors[10]. One of the Methods of detecting ASD is using the Gene Expression Analysis, as ASD is known to have a strong genetic component.

In order to measure this strong genetic component, tests were made to measure it on twins; early twin studies suggested that heritability of ASD was 80-90% [4].

2. Research Methodology:

2.1 Preliminaries

This section is an explanation of the used algorithms, their definitions and their use in the proposed approach

2.1.1 Association Rules:

Association rule learning is a rule-based machine learning method for knowing relations between variables in large databases. It finds strong rules discovered in databases using some measures, rules are created by analyzing data for frequent if/then patterns and using the support and confidence to identify the most important relationships/rules. They find all sets of items (itemsets) that have supported greater than the minimum support and then using the large itemsets to generate the desired rules that have ***confidence*** greater than the minimum confidence. A typical and widely used example of association rules application is market basket analysis. [13]Where, the support determines how many times the rule is applicable to the used dataset, and the confidence determines how many times item Y appears in transactions that have the item X, so it can be defined using the following equation [14]:

Support: (1)

$$s((X \rightarrow Y)) = \frac{\sigma(X \cup Y)}{N}$$

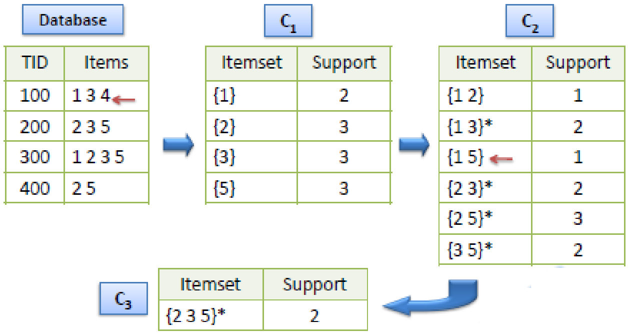
Confidence (2)

$$c((X \rightarrow Y)) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

2.1.2 Apriori Algorithm

Apriori is an influential algorithm for mining frequent itemsets for boolean association rules. While the following pseudo code provides the structure of the main Apriori Algorithm,

```
ProcedureApriori (T, minSupport) { //T is the database and min_support is the minimum support}
        Cκ: Candidate itemset of size K
        Lκ: Frequent itemset of size K
        L1:{Frequent items};
        for (k=1; Lκ != ø; K++) do begin
                Cκ+1 = candidates generated from Lκ;
                For each transaction t in the database do
        Increment the count of all candidates in C κ+1
        that are contained in t
        Lκ+1 = candidates in Cκ+1 with min_support
        End
Return Uκ  Lκ ;
```



Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 2}	1
{1 3}*	2
{1 5}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2

C_3

Itemset	Support
{2 3 5}*	2

Figure 1: shows an example of the Apriori algorithm.

While the key concept of the Apriori Algorithm is that any subset of a frequent itemset must be frequent. By reducing the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count, so in this database with 4 transactions, the steps are as follows:

1. Scan the database Itemsets
2. Create one frequent item Table C1 and add each itemset count in the database (ex. {1} has support of 2, {2} has the support of 3)
3. Compare the C1 Table to the minimum support count which is 2, how many itemsets satisfy the minimum support count, they all are.
4. Now we need to generate C2; the two frequent itemset table, using the table C1.
5. C2 all the Itemsets are paired together {1,2}, {1,3},{2,3}… and so on
6. How frequently do the itemset ex{1,2} appear in the database in the same transaction in D, the answer is 1
7. Repeat Step 3, Compare the C2 Table to the minimum support count which is 2, how many itemsets satisfy the minimum support count: {1,3},{2,5},{3,5}.
8. We need to find C3,all possible combinations using C2 are{1,2,3}, {2,3,5}.
9. Repeat Step 3, Compare the C3 Table to the minimum support count which is 2, how many itemsets satisfy the minimum support count.
10. Finally, the 3 Itemset that satisfies the minimum support count is {2,3,5}.

2.1.3 Bagging trees

Bagging tree is a simple and powerful ensemble method. It is a procedure that can be used to reduce the variance for those algorithms that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART).

Bagging uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression. [15]

It is used when our goal is to reduce the variance of a decision tree. The idea is to create several subsets of data from the training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. The result is an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree[16]. Given a sample training data[17]:

$(X_i,\quad y_i)$, i=1…n , For b = 1, …B (e.g., B = 100), Draw n bootstrap samples(x_i^{*b}, y_i^{*b}), i=1, . . . n, and a tree ($f^{tree,b}$) is fit on this sampled data set. and in order to classify an input $x \epsilon \mathbb{R}^p$, the most commonly predicted class is used:

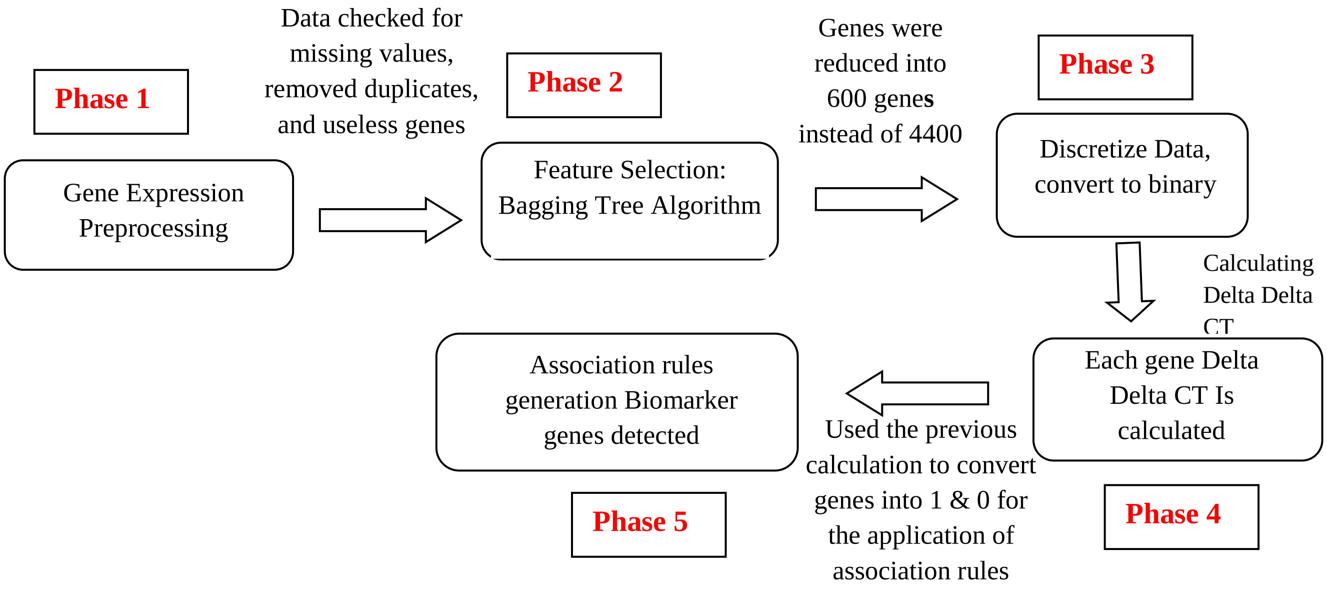
$f^{bag}(x) = argmax_{k=1,\ldots K} \sum_{b=1}^{B} 1\{f^{tree,b}(x) = k\}$

The only parameters when bagging decision trees is the number of samples and hence the number of trees to include. This can be chosen by increasing the number of trees on run after run until the accuracy begins to stop showing improvement.

2.1.4 Feature Reduction:

Feature reduction is the process of selecting a subset of only the best features in the data, it is used to remove redundant predictor variables and experimental noise, a process which mitigates the curse-of-dimensionality and small-n-large-p effects. Feature reduction is an essential step before training a machine learning model to avoid overfitting. The used data in this paper, is composed of statistical features, genes expression numbers, so in order to simplify the model, shorter the training time and reduce overfitting, this paper used the bagging tree as a feature reduction method for the enormous number of genes those exist, the dataset was more than 44,000 genes per patient, a reduction had to be done before performing the classification.

3. Proposed Model


Phase 1
Gene Expression Preprocessing
Data checked for missing values, removed duplicates, and useless genes
Phase 2
Feature Selection: Bagging Tree Algorithm
Genes were reduced into 600 genes instead of 4400
Phase 3
Discretize Data, convert to binary
Calculating Delta Delta CT
Each gene Delta Delta CT Is calculated
Phase 4
Used the previous calculation to convert genes into 1 & 0 for the application of association rules
Association rules generation Biomarker genes detected
Phase 5

The model is divided into five phases, First is Preprocessing Phase which is manual checking for duplicates and unused data, then the data is entered into the Feature Selection phase where Features number is reduced from 44000 to 700 features after that is the data discretization phase which is converting it to binary so that the Apriori can be applied and then apply a chemical genetic formula on selected features, and finally Association Rules Generation and genes detection.

4. Experimental Results

This section discusses the Dataset used and the experiment performed.

4.1 Dataset:

Microarrays: DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously, the Microarray has become an indispensable tool in biomedical research. But it also produces data with many sources of noise. A number of preprocessing prints are therefore necessary to convert the raw data, usually in the form of hybridization images, to measures of biological meaning that can be used in further statistical analysis.

Gene Expression:

Gene expression is the conversion of the DNA sequences into mRNA sequences by transcription then translated into amino acid sequences called proteins. The (GSE26415) Microarray data series, Published online by *Dr. Yuki Kuwano* in the *Gene Expression Omnibus database website* NCBI Data Repository [18].

Using DNA microarray we examined gene expression profiling in peripheral blood from 21 young adults with autism spectrum disorder (ASD) and healthy mothers having children with ASD, and 63 controls, the number of genes per person would count more than 44,000 genes. The proposed model differentiates between the control genes and ASD genes to detect the disease genes.

4.1 Preprocessing:

Although microarray dataset contains a large number of genes, a part of genes are typically excluded during the expression profiling. This process, i.e. Gene Filtering, is aimed at removing the undesirable-genes that contain outliers and too much missing expression values, and that does not exhibit variability across tissue samples so in this phase, the dataset was manually checked for missing values and some unnecessary genes and duplicated data removed

4.2 Feature Reduction:

Bagging Trees Algorithm: The bagging tree algorithm was used to help us make the decision of useless data as it works as classification and regression, it will not cause overfitting to the model, and it can handle large datasets like ours, also the more generated trees, the more accurate the results are it takes parameters the number of trees, training data, and class for each data item. In the Bagging Tree, each tree in the ensemble is built from a sample drawn with replacement from the training set. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree. As a result, the bias of the tree increases slightly, but due to the averaging of less correlated trees, its variance decreases, resulting in an overall better model[19].

The tree selects values according to its influence on the decision of whether the value is valid or not based on the previous bagging tree training set by user, then it takes each value (Node) and create decision tree, then it creates multiple decision trees according to the specified size by user If the 1s count is >0s for this values then the decision is 1, So in this Figure (2), if the value for the selected feature (x30893) Node A is less than the value 2.39326 which is the average value set by the training for this specific gene then tree will decide to give it a 0 and create another decision tree and select randomly another feature which is x35957 and check its specific value 45.573 and make the 0 or 1 decision

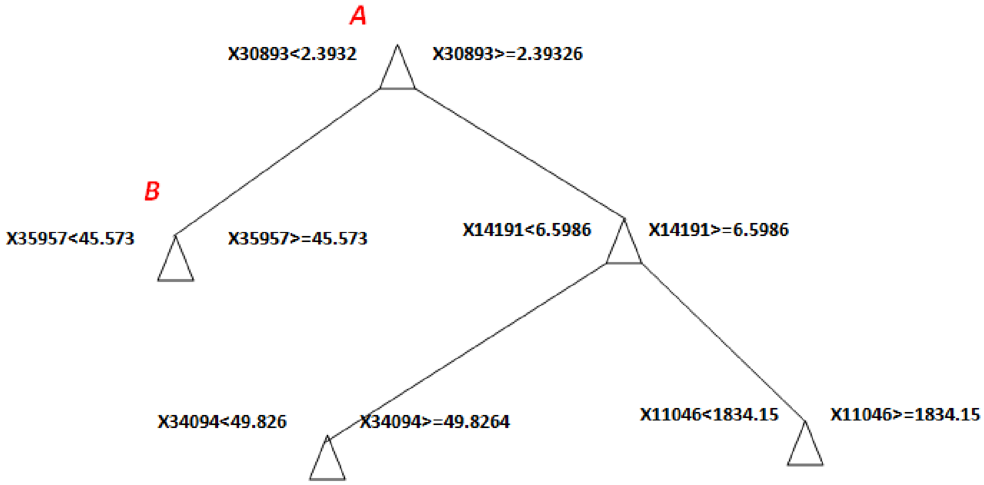


Figure 2: Output Bagging tree Sample

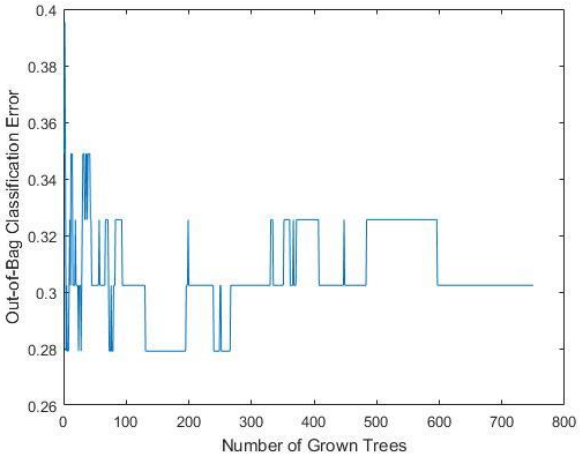


Figure 3: Out of Bag Error

The *out of bag classification* (OOB) error is the unselected features by the tree used to test the performance of the tree, the above figure shows that it improves when we increase the number of trees from the first tree to the last tree, OOB shows the variance error to show how noisy the data is and how we need to remove irrelevant genes.

The Bagging tree algorithm Results: The output of the input number of genes: 44,000a number750 trees were generated, and the number of genes we reduced to 627 genes.

3.2.1.3. Data Discretization

The data had to be discretized; convert it to binary, as in the Apriori market basket analysis approach, the item is either purchased or not, so it takes only binary.

	Gene1	Gene2	Gene3	Gene4	Gene5
Patient1	1	0	1	0	1
Patient2	0	0	0	1	1
Patient3	1	1	1	1	0
Patient4	1	0	0	1	1

Table 1: Discretization Process Example

Table 1 shows an example of the dataset after discretization, Overexpressed genes, and under-expressed genes are given the value 1 while normal range genes are given the value of 0. Using the 63 control patients, we first had to calculate the delta delta CT, so we calculated the average per gene from the reduced control genes from the bagging tree output and subtracted it from each gene in ASD:

ΔΔCt = ΔCt (treated sample) – ΔCt (untreated sample)[11]

A threshold was selected 0, genes that are with a negative value were considered down-regulated so we gave them a 1, and genes with positive value were considered 0.

3.2.1.4. Association Rules Result:

The reduced gene was now in a binary state and was used to create the rules that eventually show us the ASD biomarker genes or group of genes, as it could be a combination of two genes or three that causes the disease, the following is the output of the final Apriori code:

Gene 1	**Gene 2**	**Support**	**Confidence**
A_23_P351328	A_23_P83175	0.9	0.83
A_23_P318039	A_23_P139825	0.9	0.83

A_23_P59470	A_24_P387609	0.9	0.83
A_24_P63827	A_24_P919920	0.9	0.83

Table 2: Output Genes support & confidence

4. Result Discussion:

The generation of association rules is dependent on the minimum support and confidence threshold values: the higher these values, the smaller the number of rules generated, holding other parameters constant [20], as shown in the Table 2the validation of the rules reached support 90% and confidence 83%.

After the rules were generated, an equation was developed to remove the redundant rules and/or the redundant itemsets per rules, as they can be displayed many times with different antecedent and consequence order, then we displayed the output rules in the above table with a maximum itemsets in the rules are two, let us explain first rule meaning for example:

A_23_P351328	A_23_P83175	0.9	0.83

The right hand sided – consequent gene referred to as (A_23_P83175) is 90% (support) under-expressed when the Left hand sided gene (A_23_P351328) is under-expressed in ASD patient and this rule is 83% of the time correct.

Finally, the Biomarker genes discovered in this paper are:

ID	**Gene Symbol**	**Gene Name**
A_23_P351328	C2orf53	chromosome 2 open reading frame 53
A_23_P83175	PTPLAD2	protein tyrosine phosphatase
A_23_P318039	RQCD1	RCD1 required for cell differentiation1 homolog (S. pombe)
A_23_P139825	YAF2	YY1 associated factor 2
A_23_P59470	ZNF467	zinc finger protein 467
A_24_P387609	ISCA1	iron-sulfur cluster assembly 1 homolog (S. cerevisiae)
A_24_P63827	DNAJB6	DnaJ (Hsp40) homolog, subfamily B, member 6
A_24_P919920	ARV1	ARV1 homolog (S. cerevisiae)

Table 3: ASD Biomarker Genes Discovered

This paper discovered genes are chromosome 2 open reading frame 53, protein tyrosine phosphatase RCD1 required for cell differentiation1 homolog (S. pombe), YY1 associated factor 2, zinc finger protein 467, iron-sulfur cluster assembly 1 homolog (S. cerevisiae, DnaJ (Hsp40) homolog, subfamily B, member 6 and ARV1 homolog (S. cerevisiae), the genes were the product of classification association rules with support 90% and confidence 83%.

4.1. Previous Researches Results*:*

List of genes discovered by the paper [6] which used the same Genes Dataset as this paper:

Gene Symbol	Gene Name	Methodology used
HSF2	Heat shock transcription factor	Performed prediction using machine learning algorithms such as support vector machine, K-nearest neighbor and linear discriminant analysis and using "set.seed" function in R, randomly divided data into training set(13 ASD & 13 control) and test set (8 ASD & 8 control)
MIER2	Mesoderm induction early response 1, family member 2	
TC2N	Tandem C2 domains, nuclear	
NPM1	Nucleophosmin (nucleolar phosphoproein B23, numatrin)	
PKM	Pyruvate kinase, muscle	
ARHGAP15	Rho GTPase activating protein	
RFX1	Regulatory factor X, 1	
MRPS31	Mitochondrial ribosomal protein S31	
C12orf29	Chromosome 12 open reading frame 29	
JADE2	Jade family PHD finger 2	
ACKR3	Atypical chemokine receptor 3	
TAPT1-AS1	TAPT1 antisense RNA 1	
TMEM41B	Transmembrane protein 41B	

Table 4: Previous Researches discovery

Other similar studies [4, 5, 20, 21, 22, 23, 24] Biomarker genes symbols were:

Gene Symbol	Gene Name	Methodology used
SHANK3	SH3 and multiple ankyrin repeat domains 3	Paper[4] Developed a complementary machine-learning approach based on a human brain-specific gene network to present a genome-wide prediction of autism risk genes, Support vector machine classifier was trained to identify network patterns and differentiate ASD genes That would tell ASD subject Then it would produce ranked list of ASD candidate genes
PTCHD1	the patched domain containing 1	
DEGs	degenerative spermatocyte homolog 1, lipid desaturase (Drosophila)	
NRCAM	neuronal cell adhesion molecule	

TRIM22	tripartite motif containing 22	Paper [5] is a study that reviews more than 30 twin studies of autism spectrum disorders (ASD) and autistic traits published in the last decade that has contributed to this endeavor. These twin studies have reported on the heritability of autism spectrum disorders and autistic traits in different populations and using different measurement and age groups.
FAM107A	family with sequence similarity 107, member	
SDF-1	stromal cell-derived factor 1	
ESR1	estrogen receptor 1	
IGFBP5	insulin-like growth factor binding protein 5	
LAMA2	laminin, alpha 2	Paper [21] used the Fisher Discriminant Analysis (FDA), which achieves an optimal linear separability using a typically small set of latent variables that are linear combinations of the original variable set., Kernel FDA (KFDA), exist which can take nonlinear relationships into account for classification, regression techniques include partial least squares (PLS) and its nonlinear counterpart kernel PLS (KPLS)]. Using FDA for classification and KPLS for regression allow multivariate interactions to a surface, which are often hidden when only univariate analysis is considered
RNF144B	ring finger protein 144B	
ASXL3	additional sex combs like 3 (Drosophila)	
NLGN4X	neuroligin 4, X-linked	
DHT	dehydrogenase E1 and transketolase domain containing 1	
CXCR4	chemokine (C-X-C motif) receptor 4	
TAOK2	TAO kinase 2	Papers [22,23,24] are all Medical publications that used medical procedures and laboratory genetic analysis to get the resulted biomarker genes.
ESR2	estrogen receptor 2 (ER beta)	

Table 5: Previous Researches discovery

The findings of all the research papers in the field of genetic biomarkers for ASD were different from each other, none actually had any common results, which made the results validation a hard process and makes this field still an open research area for more discussion and validation.

5. Conclusion

In this paper the problem was detecting ASD in earlier child years than other methods usually do as listed previously like CARS Scale, Eye Tracking and Brain MRI, the proposed model was used in the detection using gene expression as ASD is known to have strong genetic component, our results showed different genes responsible for said disease as shown in the above table.

The proposed approach could also be used in the detection of other various diseases, diseases that are proven to have a strong genetic signature, then the model shall be efficient and will generate the rules containing genes that are relevant to said disease. A bigger dataset will be even more efficient, also using the approach in physical disease datasets will be easily collected than mental and psychological disorders. As they require a large number of licenses and approvals that are very hard for the patient's family to give out. Another improvement in the algorithm could be by using an FP growth technique and Apriori algorithm combined, they shall form a better technique that would generate less redundant itemsets and sub-itemsets thus increase rules generation.

6. References

1. CDC Center for Disease Control & Prevention, https://www.cdc.gov/ncbddd/autism/facts.html
2. Autism Reading Room, http://readingroom.mindspec.org/?p=7160
3. G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley, "Randomized, Controlled Trial of an Intervention for Toddlers With Autism: The Early Start Denver Model," PEDIATRICS, vol. 125, no. 1, pp. e17–e23, Nov. 2009.
4. A. Ronald and R. A. Hoekstra, "Autism spectrum disorders and autistic traits: A decade of new twin studies," American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, vol. 156, no. 3, pp. 255–274, Jan. 2011.
5. A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder," Nature Neuroscience, vol. 19, no. 11, pp. 1454–1462, Aug. 2016.
6. S. H. Kim, I. B. Kim, D. H. Oh, and D. H. Ahn, "Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning," European Neuropsychopharmacology, vol. 27, p. S1090, Oct. 2017.
7. W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," Autism Research, vol. 9, no. 8, pp. 888–898, Apr. 2016.
8. N. Yahata, J. Morimoto, R. Hashimoto, G. Lisi, K. Shibata, Y. Kawakubo, H. Kuwabara, M. Kuroda, T. Yamada, F. Megumi, H. Imamizu, J. E. N áñezSr, H. Takahashi, Y. Okamoto, K. Kasai, N. Kato, Y. Sasaki, T. Watanabe, and M. Kawato, "A small number of abnormal brain connections predict adult autism spectrum disorder," Nature Communications, vol. 7, p. 11254, Apr. 2016.
9. K. Pierce, D. Conant, R. Hazin, R. Stoner, and J. Desmond, "Preference for Geometric Patterns Early in Life as a Risk Factor for Autism, "Archives of General Psychiatry, vol. 68, no.1, p.101-109,Jan. 2011.
10. Y.-H. Nah, R. L. Young, and N. Brewer, "Using the Autism Detection in Early Childhood (ADEC) and Childhood Autism Rating Scales (CARS) to Predict Long Term Outcomes in Children with

Autism Spectrum Disorders," Journal of Autism and Developmental Disorders, vol. 44, pp. 2301–2310, Mar.2014

11. Analyzing your QRT-PCR Data-The Comparative CT Method (ΔΔCT Method), HTTP:// www.science.smith.edu
12. N. Yahata, J. Morimoto, R. Hashimoto, G. Lisi, K. Shibata, Y. Kawakubo, H. Kuwabara, M. Kuroda, T. Yamada, F. Megumi, H. Imamizu, J. E. N áñezSr, H. Takahashi, Y. Okamoto, K. Kasai, N. Kato, Y. Sasaki, T. Watanabe, and M. Kawato, "A small number of abnormal brain connections predict adult autism spectrum disorder," Nature Communications, vol. 7, p. 11254, Apr. 2016.
13. M. Al-Maolegi and B. Arkok, "An Improved Apriori Algorithm For Association Rules," International Journal on Natural Language Computing, vol. 3, no. 1, pp. 21–29, Feb. 2014.
14. Association Analysis: Basic concepts and Algorithms
15. https://blog.statsbot.co/ensemble-learning-d1dcd548e936
16. https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9
17. http://www.stat.cmu.edu/~ryantibs/datamining/lectures/24-bag.pdf
18. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26415
19. https://blog.statsbot.co/ensemble-learning-d1dcd548e936
20. Y. Kuwano, Y. Kamio, T. Kawai, S. Katsuura, N. Inada, A.Takaki, and K. Rokutan, "Autism Associated Gene Expression in Peripheral Leucocytes Commonly Observed between Subjects with Autism and Healthy Women Having Autistic Children," PLoS ONE, vol. 6, p. e24723, Sep. 2011.
21. Daniel P. Howsmon1,2, Uwe Kruger3, Stepan Melnyk4, S. Jill James4, Juergen Hahn "Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation" 1,2,3 2017
22. , Woodbury-Smith M, Scherer SW "Progress in the genetics of autism spectrum disorder" 2018 May; Epub 2018 Mar 25.
23. . Kara T1, Akaltun İ2, Cakmakoglu B3, Kaya İ4, Zoroğlu S5" An Investigation of SDF1/CXCR4 Gene Polymorphisms in Autism Spectrum Disorder: A Family-Based Study" 2018 Mar;15(3):300-305.
24. Melanie Richter, Nadeem Murtaza, Froylan Calderon de Anda "Altered TAOK2 activity causes autism-related neurodevelopmental and cognitive abnormalities through RhoA signaling" *Molecular Psychiatry* (2018)