VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2022, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 10, Number 1, January-June, 2022 *pp:42-56*

Identifying Disease Genes based on Machine Learning approaches for Classification

Rahu Sikander[1*], **Ali Ghulam**[2], **Mujeeb -ur- Rehman**[3]

[1]Department of School of Computer Science and Technology, Xidian University, China; [2]Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan; [3]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China

Keywords: *Autism spectrum disorder (ASD), Machine Learning, Disease, Genese, ASD genes.*

Journal Info:
Submitted:
May 15, 2022
Accepted:
June 16, 2022
Published:
June 30, 2022

Abstract In recent years, researchers have become increasingly interested in disease-gene assication prediction. In the postgenomic era, this is one of the toughest jobs around. It is also challenging to determine biological research since complex disorders sometimes have very varied genotypes. Machine learning methods are used widely in the identification of crawl marks, but their images depend heavily on their quantity and quality. In crawling studies, we find that the recognition of genes reconciling diseases can be improved by an machines classifier qualified in practical gene seamlessness from gene ontology (GO). In order to predict the genes of the disease, we've developed a supervised machine learning system. In the proposed pipeline, the use of autism spectrum disorder (ASD) is assessed. Similarity tests from various semantitics have been used to quantitatively measure similarity in gene function. In this paper we suggest various techniques for classifying data from one-hot encoding method. This experiment is complicated by the fact that the into training and test sets. This is generally called an algorithm evaluation divided-train-test split method. ASD is a disease associated with high health care costs and early intervention will significantly minimize these costs. ASD is a neurodevelopment disorder. Unfortunately, wait times are lengthy for an ASD diagnosis and treatments are not cheap. The economic effects of autism and an increase in ASD cases worldwide show an urgent need to establish methods of screening that are quickly enforced and efficient. A timely and affordable ASD screening is therefore imminent to help health practitioners and to let individuals know whether they will be formally diagnosed clinically. Classifiers qualified and validated for ASD and non-ASD genes work better than ASD classifiers previously reported. For instance, in order to predict new ASD genes, the complementary forest classification (CF) classification reached AUC 0.80 above the reported classification (0.73). Continuing, 73 novel ASD candidate bases can be predicted by the classifier function. Such genes enrich the central ASD syndrome, such as autism and compulsion.

***Correspondence Author Email Address:**
sikander@stu.xidian.edu.cn

1 Introduction

Most of the genetic risk factors for autism spectrum disorders (ASD) are also found in the non affected general population, making ASD a genetically complex and heterogeneous group of neurodevelopmental conditions[1]. The first objective of this study was to use resting state functional magnetic resonance imaging (rs-fmri) data to classify participants according to autism spectrum disorders (ASD) and their respective neural patterns of functional connectivity. We use a deep learning method that combines supervised and unsupervised machine learning (ML) methods[2]. Individuals with ASD display rigid repetitive movements, limited interests, lack of impulse control, speech deficits, and impaired intellectual and social skills compared to typically developing (TD) children, so this study set out to learn more about this group of polygenetic developing brain disorders[3]. Conditions that fall under the umbrella of autism spectrum disorder (ASD) are difficult to define but have underlying deficits in social interaction, communication, stereotyping, and restricted behaviour. In this investigation, we want to deduce the relationship between polygenetics and neurological disorder[4]. People with autism frequently struggle with perception and attention issues, despite the fact that these issues are not important diagnostic categories. These constellations are believed to suggest broad cortical dysfunction in the patient[5]. According to Minshew and William, autism is a chronic syndrome marked by aberrant behaviour and a lack of social reciprocity. It is a disorder of polygenic development and neurobiology. ASD is a broad category of behavioural problems with core deficiencies in communication, social interaction, stereotypes, and constrained conduct. People with autism frequently struggle with perception and attention issues, despite the fact that these issues are not important diagnostic categories. These constellations are believed to suggest broad cortical dysfunction in the patient[5]. Understanding the particular clinical process of those with autism spectrum disorder (ASD), which affects one out of every 68 children under the age of 68, is of tremendous clinical importance. Particularly, the comorbidity patterns of people with ASD have been under increasing scrutiny. Effective ASD classification has wide-ranging ramifications. Clinical course variables can assist stratify the likelihood of subsequent issues and provide details on more effective treatment trajectories. Another first step towards more strong genomic and molecular research that could result in a better understanding of the causes of ASD is segmenting the patient population into more homogeneous subgroups[6]. Autism spectrum disorder (ASD) is a group of developmental disorders that can lead to difficulties in cognitive health, social communication and social interaction. ASD affects about 1% of the world's population. It is usually diagnosed in childhood. Autism affects one in 68 children in the United States. Since 2000, the prevalence of ASD among American children has increased by 120 percent. Today, ASD is one of the fastest growing developmental disorders, causing approximately $250 billion in health care costs annually in the United States [centers for disease control and prevention, 2014[7]. Early identification of ASD in young children is often delayed due to a combination of child and healthcare factors. At the child level, the presence of early markers in the pathogenesis pipeline and the pattern of variability may hinder early recognition. In health care, early screening is further hampered by the lack of knowledge and expertise of many primary health care providers in ASD and the lack of infrastructure to handle the increase in referrals. This study examines the usefulness of automated policies in validating initial parental concerns in web-based sources. In the long run, such systems could be programmed to prompt parents to seek professional advice when their concerns indicate a risk [8]. The aim of the study was to investigate the relationship between autoimmune diseases in pregnant mothers, history of allergies and asthma, and subsequent diagnoses of childhood autism spectrum disorder (ASD). Diagnoses of autoimmune or allergic diseases recorded by physicians may affect the brain development of the fetus, leading to autism in some genetically susceptible individuals[9, 10].

The first aim of this study was to use resting-state functional magnetic resonance imaging (rs-fMRI) data to classify participants by autism spectrum disorders (ASD) and their respective functional connectivity neural patterns. We use a deep learning method that combines the methods of supervised and unsupervised machine learning (ML). Psychological disorders such as autism spectrum disorder (ASD) are complicated to diagnose exotic diseases especially in children. Current clinical assessment procedures are entirely based on behavioral symptom findings (DSM-5/icd-10) and may be vulnerable to misdiagnosis [11]. Kids with autism spectrum disorder (ASD) and 13 kids with low functioning non-asd disorders. Typically,

ASD is characterized by impaired social interaction, narrow interests, and repetitive behavior, with wide variations in expressions and severities. Brain imaging studies show several findings suggest that ASD is correlated with an unusually complex and distributed mode, which is a daunting task in recognizing typical and neuroimaging maps. In this case, our goal was to classify the remaining trends of ASD-related functional brain imaging anomalies and to check their efficacy in the person classification[12]. Studies of infants, adolescents, and young adults with ASD showed a lot of health problems. Factors of risk which can lead to unnecessary morbidity. Poor eating habits are among the health risk factors observed in young people with ASD. Many people with autism are taking multiple psychotropic medications for obesity and limited physical activity [13]. Most morphological studies at ASD have used morphometric (VBM) based on voxels, which tests changes in voxel orientation and volume of white matter throughout the brain. Such VBM-based experiments calculated the differing amount of gray and white matter. The intrinsic topology of the neocortex, however, is that of 2-d slices with a highly folded and curved geometry, and VBM cannot measure this topology directly. Morphometric surface (SBM) focuses on the estimation of topographic cortical dimensions, which can supply information to Neuroimaging 50 (2010) 589-599 complements the information provided by VBM. SBM may derive characteristics such as area thickness and area surface area gray matter [14].

Despite the existence of robust diagnostic tools for ASD, physicians still use their daily practice with a variety of tools and procedures. Classification of ASD by the "gold standard" method used in clinical settings is often not possible for large-scale or population-based research. Alternatively, many epidemiological studies often rely on current "administrative" terms for ASD classification: the international disease classification, the 9th amendment (icd-9) billing code, categories of special education, or eligibility for disability benefits and autism programs (such as Medicaid). The outlets that use these classifications vary widely in the United States, and since their primary purpose is to ensure that people are provided with adequate care rather than classifying disabilities, not all persons that meet the ASD requirements are consistently defined in those schemes [15]. This analysis discusses the latest work pertaining to models of ASD disease that use IPsec as a source of stem cells to produce in vitro nerve cells. It is clear from the findings summarized in this paper that ASD disease modeling is already advancing our understanding of the disease etiology. Furthermore, the method offers an unparalleled opportunity to manipulate neural ASD networks in a controlled environment to test approaches to restore altered neuron phenotypes. Such results might also allow us to understand other neurodevelopmental conditions better [16]. A representative study in Sweden has been investigating mortality in ASD patients as compared to a control group to date. The study examined the all-cause mortality of ASD patients registered in the Swedish national registry of patients from 1987 to 2009, based on population. The findings showed that people with ASD died from injuries (i.e., asphyxia, suffocation, drowning), epilepsy, birth defects, malignant tumors, respiratory and circulatory disorders, and suicide compared to people without ASD at a younger age (20 years younger than their era) [17]. ASD is known to be a developmental brain disorder affecting some communication and social behavior. There are many methods to treat ASD. Clinical diagnostic purposes, for example, include revamped ados-r, ADI, etc. Clinical testing approaches have shown comparative efficacy in screening for ASD-related cases. For example, in several different experimental studies, Ados-r and ADI obtained good sensitivity and specificity results. Furthermore, both approaches show acceptable high reliability and effectiveness performance. Cancerlectins are lectins that are strongly associated with particular types of proteins that start the survival, development, metastasis, and spread of cancer cells. In the post-genomic era, differentiating a protein based on its functioning is still a challenging task[18]. Autism Spectrum Disorder (ASD) is a nervous system developmental disorder. The autism spectrum phenotype is well described but is poorly understood for its etiology and pathogenesis. According to the results, genetic and environmental risk factors are the principal causes of autism. Most ASD symptoms usually begin around the age of 2 so early diagnosis is needed. It is accepted that current clinical approaches do not make a proper distinction between patients and healthy controls (HCS). Yet avoiding the difficulty of identifying abnormal brain areas in people with ASD to avoid these deficits is not difficult enough, so machine learning was applied in the field of neuroimaging. This is an important way of extracting information from neuroimaging data and predicting more possible disease changes [19]. Although an increasing number of reports reliably affirm evidence supporting early over the wiring in human disease-pathway association

research is a recurrent area of interest for the biomedical community. Finding the processes or connections between diseases and pathways can be aided by this linkage. Despite decades of research in this field, the accuracy of disease identification has been less than ideal. In order to forecast disease-pathway connections, this study suggests a computer model. To implement early intervention approaches more effectively [20].

Recently, non-invasive brain imaging approaches have been combined with advanced machine learning (ML) technology-based image analysis to provide automated detection of neuropsychiatric disorders, thus improving or confirming their diagnosis. A diagnostic tool for ASD diagnosis has been developed over the past decade, based on the structural magnetic resonance imaging (SMRI) of the brain (see the "gold standard" diagnosis of neurodevelopmental disorders is costly, subjective and time-consuming. Relying on a multidisciplinary team of neurologists, pediatricians, pathologists, and therapists is vulnerable to explanatory bias, frequently requiring years of continuous interviews and behavioral evaluations. Autism spectrum disorder (ASD), for example, affects one in sixty-eight individuals in the United States. Early diagnosis and intervention were not widely available or implemented due to the prevalence of ASD and limited health services. Influential people are often ignored in their infancy. So quantitative and unbiased diagnostic tools have been greatly needed for decades, but little progress has been made [21]. Recent advances in wearable sensor technology, in particular the production of IMU sensors, have provided an important forum for remote monitoring of patients with movement disorders such as Parkinson's disease (PD) and autism spectrum disorders (ASD). The IMU contains internal accelerometers, gyroscopes, and sensors for the magnetometer, which measure angular velocity and linear acceleration of body parts during motion. IMU has become the most popular device for human motion recognition and irregular motion detection, thanks to its small size, high portability and lightweight. IMU not only offers the possibility of automatically measuring complex symptoms and phenotypes, particularly in psychiatric clinical studies but also enables carers to monitor disease progression and intervention quality more regularly than current clinical practice [22]. In this research, we hypothesized that machine-learning techniques could be used to predict disease-related genes throughout contrast to current scientific and technological approaches in functional similarities only. Performance was increased. We also checked that the USES classification is the same form of evaluation used by Krishnan et al. Five layers of cross-validation also validate the classifier. The ASD matrix and non-psychiatric genes are classified as a classifier, classifying and teaching genes only, using various machine learning approaches to functional similarity. The interface is greater than the approaches already in use. This suggests that GO functional similitude is a tool for ASD gene prediction. Machine learning Methods for semantic similarity educated and validated can therefore reduce essentially the genetic significance of ASD. A highly reliable ASD gene-based classifier was also developed to boost its effectiveness so that this can be used for the prediction of ASD genes. The KNIME workflow of the proposed solution is also shown fully automatic and customizable. It can be used to diagnose any other kind of gene disorder.

2 Methods and Data Source

2.1 Data Source

This real-world dataset was discovered on and includes information on 464. There are 7 gene annotation features in the data. Certain characteristics, such as gene status and gene score level, are continuous. Some are categorical, such as syndromic and report type. The goal is to forecast the seventh feature—the presence of autism spectrum disorder (ASD) disease in a patient and non-associated autism spectrum disorder—using the first seven features (ASD). The fact that this experiment uses training and test sets makes it more challenging. Because even a basic (by modern standards) neural network contains more degrees of freedom than there are data points in our sets, we must exercise extreme caution.

2.2 Data Overview of the proposed method

In this research, we developed a machine-based approach for identifying genes for disease. The general custody of the system suggested is shown in Figure 1. Functional matrices for disease-related genes were designed for any set of functions (step-1) For each set. The gene expression profiles, protein-protein interaction network or GO will decide the functional resemblance between genes. In the step-2 in the framework we extracted features based on One-hot encoded method. We then used various classification machine learning classifers and then trained with in step-3. We then deploye the model validation based on 5-fold validation method and then evalue the performance metrics with in step-4.

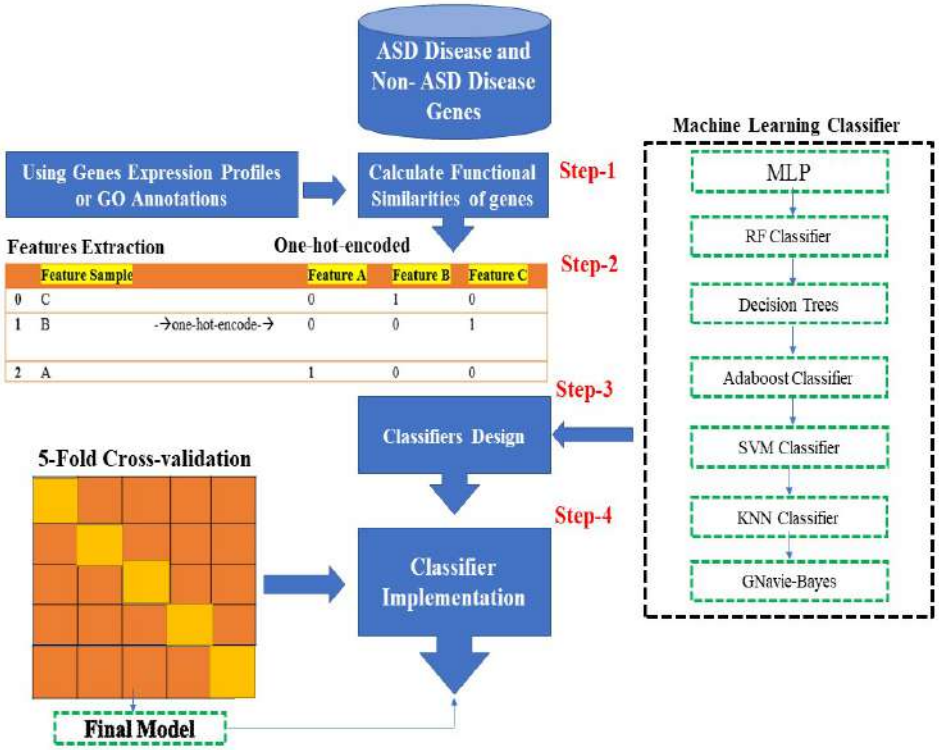

Figure 1. Flowchart of proposed method to predict disease genes.

2.3 Data Preparation

It must be cleaned, formatted and maybe even restructured before the data can be used for machine-learning algorithms as input – usually known as preprossessing. The outcome and predictive power of almost all study algorithms can be greatly helped by this preprocessing.

2.4 Correlation between different features

In the above two examples, we have used two separate representation strategies to look quickly at our ASD dataset. We used the seaborn visualization software “factor plot” module. In the first case, we used ’Swamp’ graph expressing the relationship between several different characteristics of the data, while in the second case, the ’box’ approach showed how different characteristics were related as shown figure 2(a).

2.5 Features extractions

First, we have to convert the dataframes from pandas to numpy arrays that scikit learn can use. Let's construct an array that extracts only the feature data in which we are concerned and another array containing the ASD disease class or non-associated ASD disease. The subjects of this study consisted of split the data into features and target label.

2.6 One-Hot-Coding

From the table below, we can see that there are many non numerical features, including country of residence, ethnicity and so on for each record. Learning algorithms usually anticipate numerical input, which involves converting non-numerical features (called category variables). The one-hot encoding system is a popular method for converting categorical variables. For of type of non-numeric, one-hot encoding produces a "dummy" vector. Assume that some feature contains three entries possible: A, B, or C, for example. This function is then encoded into function A, Feature B and Feature C as shown in Table 1.

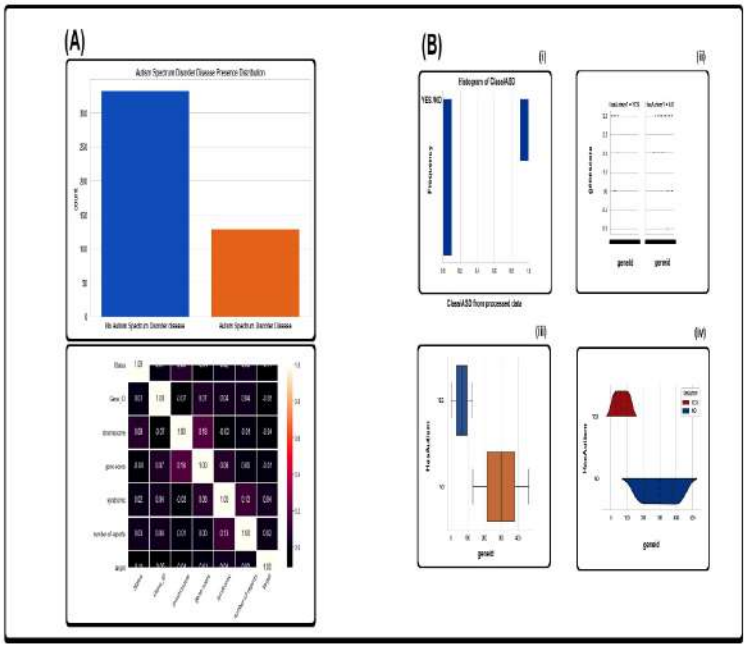

Figure 2. Distribution positive and negtive sets, Correlation between different features and visualizations

Table 1. Feature Sample

Feature Samples		Feature A	Feature B	Feature C
0 C		0	1	0
1 B	->one-hot-encode->	0	0	1
2 A		1	0	0

Therefore, we have to transform the non-numeric target mark 'Class / ASD' in numeric values for the learning algorithm, as with non-numeric functionality. Seeing that there are only two separate categories ('YES' and 'NO' to' Class / ASD 'for this label), we can stop using a one-hot encoding and represent the two categories as 0 and 1 as shown in figure 2(b).

$$x = (x : 1 if x ==' YES' else 0) \quad (1)$$

2.7 Histogram of Class/ASD

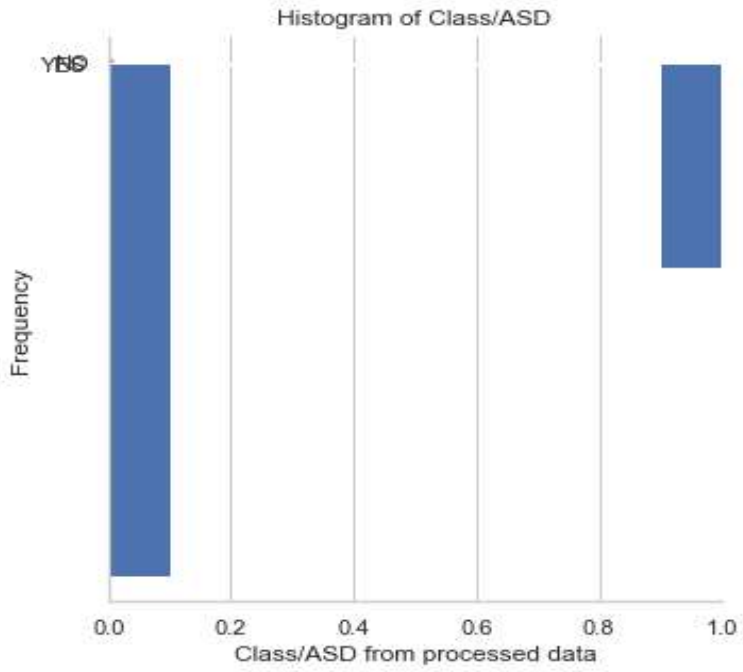

Figure 3. Correlation Histogram of Class/ ASD

2.8 Split data train and test set

All categorical variables have now been transformed and all numerical characteristics have been normalized. As always, we will split the data into training and test sets (both features and their labels). For training and testing, 80 percent of the data will be used and 20 percent as shown in table 2.

Table 2. Train and test datasets

Preprocess datasets	Normalized Data
Total number of records: 463 Individuals diagonised with ASD: Individuals not diagonised with ASD: 333	Training set has 364 samples. 130 Testing set has 92 samples.

2.9 Supervised Machine learning models Implementation

The following supervised models of learning that are available at present in our proposed method. we have used two separate representation strategies to look quickly at our ASD dataset. We used the seaborn visualisation software factorplot module. In the first case, I used 'Swamp' graph expressing the relationship between several different characteristics of the data, while in the second case, the 'box' approach showed how different characteristics were related. We used various supervised models of classification in this study, including machine learning, such as Decision Trees[], Random Forest[16], Support Vector

Machines (SVM)[15], K-Nearest Neighbors (KNeighbors)[],Gaussian Naive Bayes (GaussianNB)[14], Logistic Regression Linear Discriminant Analysis (LDA)[] and Quadratic Discriminant Analysis (QDA)[]. Overseen techniques of machine learning for the detection of diseases have been widely used. Le et al. [17] can find some sort of comparison of category-based approaches.

2.9.1 Random Forest models Implementation

NB is a simple gas probability classification, based on the theorem Bayes, assuming that each of the stored range variables is firmly (naively) independent. The current basic Bayesian grade level is technically done, obviously, even if independence deficiency is assumed by a conditional likelihood model, which uses the Bayes theorem to estimate the distribution of gas probability for every spot, and then uses the Spot Time Test for the actual savings test class. e1071r is used by NB. When using NB, do not draw.

2.9.2 Support Vector models Implementation

SVM is a heuristic method of optimization, which tends to findobserver co-orders. A support vector for the determination of hyper-decision planes that better separate these two groups (more or less) has the maximum range in n-dimensional space where n is the number of modelling features. SVM can be used by extension as a feature of the Kernel linear or non-linear process. For non-linear SVM, the kernel is replaced by the dot product of linear SVM. The mapping of data in a broad feature space is crucial. In this analysis we have been predicting disease genes using linear SVM and radial kernel-based SVM. R-package e1071 (version 1.6.8) [19] was used for the lineal and the radial SVM. The weight of equity describes positive and negative categories. Price and epsilon parameters are used for value 1 and 0.001. The filter is set to 0.02 for radial SVM.

2.9.3 Naivey Bayes models Implementation

NB is a simple gas probability classification, based on the theorem Bayes, assuming that each of the stored range variables is firmly (naively) independent. The current basic Bayesian grade level is technically done, obviously, even if independence deficiency is assumed by a conditional likelihood model, which uses the Bayes theorem to estimate the distribution of gas probability for every spot, and then uses the Spot Time Test for the actual savings test class. e1071r is used by NB. When using NB, do not draw.

3 Performance Evaluation of the classifiers

Two distinct assessment approaches, stratification five-times cross-validation and retaining of selective five-fold stratified cross-validation adapted from Krishnan et al., all classifiers were evaluated. Study. - Research. The following steps are taken by layered fivefold.

1. Divide the data set into five equal folds, whose class probability is similar to the original data set
2. Training on four randomly selected fold-up training classifier groups (training)
3. Use the remaining folding traces (test set) to test the singer training classifier.
4. This process was repeated five times so that the industry would use this kind of folding traces as a ranging test set.

In a restricted layered five-fold cross-validation, use a sub-classified to test the set Tea from the test set (step three), named Keep similar to Krishnan et al., the quota estimates are used to compare the effects of all the produced classifers in the area under the recipient-operator curve (AUC). Sergey The self-similarity rating score is skewed. Semantic consistency is squared during cross-validation. The matrix is folded by 50%. The gene chosen in the test set as an example has been excluded from Test and Training Center Functions. I went like this abroad, for instance a test set of five different genes: g1... Semantic resemblance is a series of 5 — 5 centimetres. G1... G1... G3 uses training set examples, and in test sets g4 and g5 are not valid, so they are used as the feature of training set. The matrix capacity reduced by 3 bases g1... g3 to 3x3 limit. So what biases have been triggered abroad, the instance test set, and the workout set feature are omitted, which decreases the test and training set functionality.

3.1 Class imbalance effects

3.1.1 Class imbalance effect

Under-sampling is used to deal with positive and negative imbalances. The adverse situation was observed. During classifier design, most categories (negatives) are Insufficient samples were sampled in each of the five cross-validation training folds. To this end, PercPos Use the method in the non-balanced R package and sets the perc value to 30 to get a subset 30% positive and 70% negative. The classifier is built for fivefold cross-validation Repeat 20 times on the under-sampled dataset and report the average AUC value.

4 Method evaluation and performance

We investigated various ML models for this proposed in this study. The scikit-learn and tensorflow libraries for Python was our primary resources. Yet, it should be noted right away that we shouldn't expect a model to be 100% accurate. In fact, ourporposed model performs exceptionally well, as its success is typically the product of a wide parameter space. We considered something pretty comforting: the majority of classifers reach an accuracy of.885. Its stability across a wide range of model types points to a bayes Error of probably 8–10%.

4.1 Metrics measure

We may use F-beta score to calculate both precision and recall:

$$F_{=(1+^2).(precion.recall)/((^2.percision)+recall)}(2)$$

In hence, when = 0.5, more emphasis is placed on precision. This is called the F. 0.5 Score (F-score for simplicity).

$$Accuracy = 1/N_(i = 1)^N |X_iY_i|/(X_iY_i) \quad (3)$$

$$Sensitivity = TP/TP + FN \quad (4)$$

$$pecificity = TN/FP + TN \quad (5)$$

$$F - measure = 2(RecallPrecision/Recall + Precision) \quad (6)$$

$$Precision = TP/TP + FP \quad (7)$$

Accuracy tests how much the classifier forecasts accurately. It is the ratio of number of correct predictions to overall predictions (number of evaluation data points). The precision metric tells us what proportion of the spam messages we reported were in fact spam. It is, in other words, a ratio of true positives (words categorised as spam) to all positives (words classified as spam regardless of whether the classification was accurate). Recall (sensitivity) reveals the proportion of spam communications that we label as such. It is a ratio of true positives (words marked as spam and indeed spam) to all of the phrases that were actually spam.

5 Results

We tested master learning methods in ASD candidate genes prediction. In this review. This is why ASD genes are functionally similar and the use of semantic similitation measures calculates non-psychogenic génes. The ratings are calculated using three different semantical similitary tests.

The functional similarity of the ASD and non-psychiatric genes was calculated using, Wang, and Rel actions. For each semantine similarity matrix, four different machine learning methods (RF, NB, linear SVM, and radial (SVM) were performed, with 9 different machine learning classifier tests. Such classifiers involve

qualified and verified RF-based classifiers. All semanthropic semblance matrices are superior (Table3). The best accuracy results were obtained in different RF classifiers, despite the use of Resnik semantic score for training. The ROC(AUC) value difference is small, indicating the independence of the semantic measurement method. Unable to distinguish HD ASD genes, both linear and radial SVM. Cross analysis of HD and all non-psychological genes has been used in the assessment of five individuals.

Table 3. Result performance accuracy and ROC Auc score

Models	Accuracy Score of:	ROC AUC Score of:
Random Forest	0.9562	0.9653
Decision Trees	0.9496	0.9488
AdaBoostClassifier	0.9496	0.9488
SVM	0.9454	0.9653
K-NN	0.79787	0.8961
GNaive Bayes	0.6974	0.8831
MLP	0.7032	0.7608

The classifier was also constructed using HD + LD functional similarity and non-psychological genes. The test has been conducted by automated five times. The output of each HD + LD gene and non-psychological gene semi-simulation matrix classifier was lower than other Hd-based, linear and radial SVM classification classifications (table 3) was excluded. The Resnik semantic HD / LD gene and non-psychic gene scores have been tested. Rf classifiers are well qualified. The AUC value of the other HD + LD gene classifications was greater. Similarly, the concentrations determined using the measurements Wang, Resnik and Rel are comparable for the RF classifier trained and checked for semantitic similarities between HD+LD (Table 3).

Krishnan et al. (2005) modified a limited method of validation. To this end, the functional similarity between HD+LD gene and non-psychological gene has been established by a classifier. Nevertheless, only HD ASD and not genes have been tested during 5- fold cross-validation datasets. RF classificators which continue to have a restricted performance better than HD and HD + LD (evaluated by hierarchy of five-fold cross-validation) for each semántic similarity matrix. Although efficiency compared with the HD and HD + LD classifications has improved, ASD genes have not been categorized as linear and radio-svm which remain constrained (Table 3). The NB classifier was tested only for HD genes and the test was performed without restriction and improved performance than other classifiers. The semanticized similarity matrix is also computed with the combined Boot Mean Value (BMA). We used the best classifier to predict new ASD candidates from a new cnv-damaged gene. In order to calculate the semanticized similitude of the matrix with HD and non-psychological genes the Resnik training and test.

6 Case study dataset

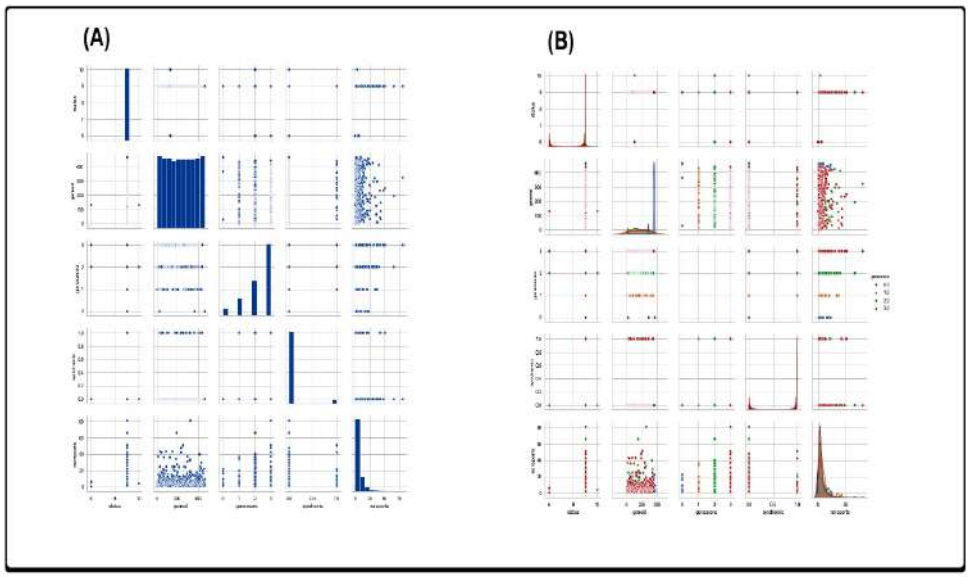

Figure 4. Autism Spectrum Disorder Disease Presence Distribution

Machine learning is the process of training a computer to statistically analyze a group of observed data and learn about possible patterns. It has recently been used to perform prediction tasks in psychology, such as the emerging field of multimodal human perception, which is focused on a computer learning voice and analyzes the human voice technology based on machine learning (e.g., emotional), Before some research was applied to autism research, based on behavior observation or brain activity. Despite that fact, in autism research, the importance and power of machine learning have not been fully discovered. While machine learning generally involves the selection of features, selection of features, model learning and prediction, many computer-based autism studies have concentrated on how to efficiently pick a subset of features from a large number of existing structured features. Scaled diagnosis to shorten diagnostic time [23]. In the case of ASD diagnosis, the classifier is required to mark the disorder and reduce the clinical score provided the picture of an individual subject. Identifying disease-related features that contribute not only to the effective image-based diagnosis of ASD but also to a better understanding of the disease is also important. Due to a large number of candidate features and the significant variations between data from multiple centers, selecting the best feature is a difficult task for multi-pattern multicore data. Relevant literature studies have adopted the approach of automated feature collection based on sparse learning [24]. To order to implement the approach suggested, it is necessary to obtain proof from the autism gene database (n= 990) (https: /gene.sfari.org/), accessed March 2018, for the genetic engineering of the Simmons Foundation Autism Recherche Project (SFARI). The SFARI gene database is a catalog of earlier studies reported with ASD-related genes. The strength of genes in the SFARI database falls into 7 different categories. Category 1 genes with certain repeatability limits are classified as genes with strongest and reproducible evidence. Categories 3 and 4 a gene with evidence for a small ASD candidate gene analysis evidence for a category 5 indirect link. There are no interactions between the six chromosomes. The database of the SFARI gene also contains genes which contribute to genetic pathology with autism signs. The symptom genes are in different categories of symptoms. Class 1 and Class 2 genes for SFARI are identified as high-reliability genes for disease (HD) (N=82), whereas Category 3 and 4 genes for SFARI are classified as Low Confidence genes for LD(N=506). From Krishnan et al. (10) is the non-psychiatric gene (N= 1189).

7 Discussion

Complex diseases have several causes and it is a challenge to identify biological recognition. For starters, ASD is a complex neurostructure. The diagnosis is based on clinical testing only because there are currently no clear biological markers. In this paper, we present an ASD prediction approach based on machine learning. Our RF classification is typically independent of and targeted only by gene weighting criteria. As far as we know, the current application of machine learning technology is the ASD gene recognition. In the research, its approach was thoroughly tested and practical applications were identified for solving biological problems. The same genes, including ASD and Non-ASD genes, were also working.

The classifier's output is likely to decrease due to weak evidence of participation in ASD, including 339 genes in group 4 of SFARI. Understanding the relation between these genes and ASD needs further research. In addition, the HD + LD ASD Genes have been trained but only HD ASD has been evaluated. A skewed estimation can however be given by determining the effectiveness of the classifications using only HD ASD genes. An number of clinical phänotypes are posed by the ASD. Only the HD ASD gene may be unrobust in genes that cause ASD or comorbidity of mild expression, reducing the prediction-classifier scope.

The benefit of this study is its independence from the weighted scheme model, based upon available evidence, to identify the genes. The classification of ASD genes by comparisons is not normal. Exclusive definition of so that genes can cause biases, which allows for weighting criteria. In this analysis, we used semantine similarity scores for genes, with a semantic similarity score close to 1 ranging from 0 to 1, which suggests high gene functionality. If the genes do not participate specifically in ASD, the target gene will have low similarities with the pathogenic gene, thereby decreasing its classification role. Therefore, no weighting requirements have been applied. The method is automated and the quality can be increased with the production of GO.

Improving the classification with semanticized similarity further demonstrates the significance of GO annotation in the prediction of disease genes. GO provides consistent and deeper data than protein-protein or gene expression. Personal information. Personal information The following drawbacks protein-protein interactions and expressions can be resolved by the practical knowledge of the GO structure, since the GO description is less context-dependent. Moreover, our study confirms that the machine learning approach can be used to predict and improve disease genes ' performance by adding more genes.

It is limited to being entirely dependent on the annotation resources available. The research filters genes without the GO annotation. However, this can improve in the future because over time the number of GO terms and remarks is increasing. We also provide a template for the design approach that is downloadable and simple to use. The workflow is intended to enable users to test several machine learning and functions in a similar way. This workflow allows domain specialists to reproduce methods for the detection and application of pathogenic genes to further research. Improve forecasts by parameter adjustment if appropriate. Developed The workflow incorporates all the resources needed to prevent genetic disease and gives researchers the chance to examine personal information without being posted to an external website in order to prevent any problems with their privacy.

Study of enrichment showed the enlargement of histone modification ASD predictive genes Way. It corresponds to previously reported studies of histone genes Participating in ASD. The predicted ASD genes were enriched significantly by only one approach. This could be attributed to the small number of ASD genes introduced. Greater predictions will contribute to a statistically more important pathway in the number of genes.

Kernel ASD symptoms are defined as inadequate social and communication interaction and repeated behavior. This core symptoms display, however, a broad phenotype, which illustrated phenotypic heterogeneity of ASD. A large range of the phenotypes are also shown in the RF Classification for the predicted enrichment analysis of the ASD gene function. Enriched autism and compulsive phenotypic behavior have confirmed further that the role of the classifiers for the prediction of new ASD genes is feasible. ASD indicates genes. Five genes also include the term HPO for the hyperactivity disorder with attention deficit. Hyperactivity with attention deficit is one of the most severe ASD complications. Most children with autism A hyperactivity disorder with attention deficit shows.

White RF is better than other approaches, but Naive Bayes (NB) is the equivalent of its performance. While RF is more efficient. Therefore, we don't need RF to arrive. The same applies to other disease genes involved. Linear and non-linear support vector machines, perhaps sensitive to the number of functions and documents, can not make reasonable predictions. The right data can also be SVM. It may not be possible to create this data set.

In this study, we documented the functional similarity of GO annotation evaluated genes for the detection of new ASD genes. The classification efficiency of the HD ASD gene was superior. All other classification systems and classifiers previously reported. It is not the aim of this task to calculate the best measure of semantic similarity for genetic functionality. Nevertheless, the efficiency of the classifier can be further enhanced by including more detailed semantical similarity. Therefore, the functions of lack of semanthetic gene marking can also be used in literature with text mining technology. It is also expected that future research will concentrate on binding protein-protein interactions with Semantic similarity scores, obtaining reliable ASD genes. In addition, details including gene expression and semantine route data will boost predictive power further.

8 Conclusion

More donation of hereditary needle transfer funds, including the attention-hungred hyperactivity disorder (ADHD), for the purpose of restoring the conditions in coexistence of ASD. Computer learning is a tool that removes the ASD mechanism's sentencing impact by discovering new disease genes and adorable guest apps. This study shows, however, that we are able to listen to high efficiency functions by integrating quantitative measurements of gene function similarity. In summary, we tested our hypothesis by using only the most current technology to improve the genetic prediction of complex GO annotation conditions. The hierarchical contrast and limitless cross-validation are a novel feature of the presentation. The results show that the classifier is based on a GO Statement on an affected classifier. The restricting ASD classifier can be realized that the semantic similarity score was just high with the five-times cross-validation process Confidence in ASD genes, which is roughly comparable in efficiency by training machine learning. This paper shows that machine learning approaches are a viable way of examining the genetic complexity of chronic disorders such as ASD. The results of the study can further contribute to the development of genetic tests and laboratory tests of genetic ASD risk factors.

Author Contributions

Rahu Sikander: Made an equal contribution to this article, Writing- Original draft preparation.**Ali Ghulam**: analysis of the manuscript, edited the manuscript to improve the English language and flow **Mujeeb -ur-Rehman**: instrumental in developing the strategy and the software, edited the manuscript to improve the English language and flow.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

Funding Information

FMS acknowledges the support of NSF grant CHE-1111111.

Author Information

ORCID:

Author 2 name: 0000-0001-5166-2213

References

[1] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," NeuroImage Clin., vol. 17, pp. 16–23, 2018.

[2] S. J. Sanders, "First glimpses of the neurobiology of autism spectrum disorder," Curr. Opin. Genet. Dev., vol. 33, pp. 80–92, 2015.

[3] I. Rapin and R. F. Tuchman, "Autism: definition, neurobiology, screening, diagnosis," Pediatr. Clin. North Am., vol. 55, no. 5, pp. 1129–46, viii, 2008.

[4] W. Jamal, S. Das, I.-A. Oprescu, K. Maharatna, F. Apicella, and F. Sicca, "Classification of autism spectrum disorder using supervised learning of brain connectivity measures extracted from synchrostates," J. Neural Eng., vol. 11, no. 4, p. 046019, 2014.

[5] Y. Zhou, F. Yu, and T. Duong, "Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning," PLoS One, vol. 9, no. 6, p. e90405, 2014.

[6] T. Lingren et al., "Electronic health record based algorithm to identify patients with Autism Spectrum Disorder," PLoS One, vol. 11, no. 7, p. e0159621, 2016.

[7] Y. Jin et al., "Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks: Identification of Infants at High-Risk for ASD," Hum. Brain Mapp., vol. 36, no. 12, pp. 4880–4896, 2015.

[8] A. Ben-Sasson, D. L. Robins, and E. Yom-Tov, "Risk assessment for parents who suspect their child has autism spectrum disorder: Machine learning approach," J. Med. Internet Res., vol. 20, no. 4, p. e134, 2018.

[9] L. A. Croen, J. K. Grether, C. K. Yoshida, R. Odouli, and J. Van de Water, "Maternal autoimmune diseases, asthma and allergies, and childhood autism spectrum disorders: a case-control study: A case-control study," Arch. Pediatr. Adolesc. Med., vol. 159, no. 2, pp. 151–157, 2005.

[10] K. N. Thakkar et al., "Response monitoring, repetitive behaviour and anterior cingulate abnormalities in autism spectrum disorders (ASD)," Brain, vol. 131, no. Pt 9, pp. 2464–2478, 2008.

[11] T. Eslami, V. Mirjalili, A. Fong, A. Laird, and F. Saeed, "ASD-DiagNet: A hybrid learning approach for detection of Autism Spectrum Disorder using fMRI data," arXiv [cs.LG], 2019.

[12] E. Duchesnay et al., "Feature selection and classification of imbalanced datasets: application to PET images of children with autistic spectrum disorders," Neuroimage, vol. 57, no. 3, pp. 1003–1014, 2011.

[13] L. Bishop-Fitzpatrick et al., "Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder: Health problems in decedents with autism," Autism Res., vol. 11, no. 8, pp. 1120–1128, 2018.

[14] J. D. Brooks et al., "Identifying children and youth with autism spectrum disorder in electronic medical records: Examining health system utilization and comorbidities," Autism Res., vol. 14, no. 2, pp. 400–410, 2021.

[15] M. J. Maenner, M. Yeargin-Allsopp, K. Van Naarden Braun, D. L. Christensen, and L. A. Schieve, "Development of a machine learning algorithm for the surveillance of autism spectrum disorder," PLoS One, vol. 11, no. 12, p. e0168224, 2016.

[16] P. C. B. Beltrão-Braga and A. R. Muotri, "Modeling autism spectrum disorders with human neurons," Brain Res., vol. 1656, pp. 49–54, 2017.

[17] M. B. Usta et al., "Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders," Psyc.. Clin. Psychopharmacol., vol. 29, no. 3, pp. 320–325, 2019.

[18] Adnan et al., "Deep-PCL: A deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features," Chemometr. Intell. Lab. Syst., vol. 221, no. 104484, p. 104484, 2022.

[19] X.-A. Bi, Y. Wang, Q. Shu, Q. Sun, and Q. Xu, "Classification of autism spectrum disorder using random support vector machine cluster," Front. Genet., vol. 9, p. 18, 2018.

[20] A. Ghulam, X. Lei, M. Guo, and C. Bian, "Disease-pathway association prediction based on random walks with restart and PageRank," IEEE Access, vol. 8, pp. 72021–72038, 2020.

[21] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[22] N. Mohammadian Rad, T. van Laarhoven, C. Furlanello, and E. Marchiori, "Novelty detection using deep normative modeling for IMU-based abnormal movement monitoring in Parkinson's Disease and Autism Spectrum Disorders," Sensors (Basel), vol. 18, no. 10, p. 3533, 2018.

[23] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," Autism Res., vol. 9, no. 8, pp. 888–898, 2016.

[24] J. Wang et al., "Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study: Multi-Modality Multi-Center Diagnosis for ASD," Hum. Brain Mapp., vol. 38, no. 6, pp. 3081–3097, 2017.