Modified Cuckoo Search-Support Vector Machine (MCS-SVM) Gene Selection and Classification for Autism Spectrum Disorder (ASD) Gene Expression

Dr.M. Kalaiarasu[1*], Dr.J. Anitha[2]

Abstract

Autism Spectrum Disorder (ASD) is a neuro developmental disorder characterized by weakened social skills, impaired verbal and non-verbal interaction, and repeated behavior. ASD has increased in the past few years and the root cause of the symptom cannot yet be determined. In ASD with gene expression is analyzed by classification methods. For the selection of genes in ASD, statistical philtres and a wrapper-based Geometric Binary Particle Swarm Optimization-Support Vector Machine (GBPSO-SVM) algorithm have recently been implemented. However GBPSO has provides lesser accuracy, if the dataset samples are large and it cannot directly apply to multiple output systems. To overcome this issue, Modified Cuckoo Search-Support Vector Machine (MCS-SVM) based wrapper feature selection algorithm is proposed which improves the accuracy of the classifier in ASD. This work consists of three major steps, (i) preprocessing, (ii) gene selection, and (iii) classification. Firstly, preprocessing is performed by mean or median ratios close to unity was removed from original gene dataset; based on this samples are reduced from 54,613 to 9454. Secondly, gene selection is performed by using statistical filters and wrapper algorithm. Statistical filters methods like Wilcox on Rank Sum test (WRS), Class Correlation (COR) function and Two-sample T-test (TT) were applied in parallel to a ten-fold cross validation range of the most discriminatory genes. In the wrapper algorithm, Modified Cuckoo Search (MCS) is also proposed to gene selection. This step decreases the number of genes of the dataset by removing genes. Finally, SVM classifier combined forms of gene subsets for grading. The autism microarray dataset used in the analysis was downloaded from the benchmark public repository Gene Expression Omnibus (GEO) (National Center for Biotechnology Information (NCBI)). The classification methods are measured in terms of the metrics like precision, recall, f-measure and accuracy. Proposed MCS-SVM classifier achieves highest accuracy when compared Linear Regression (LR), and GBPSO-SVM classifiers.

Key Words: Modified Cuckoo Search (MCS), Support Vector Machine (SVM), Autism Spectrum Disorder (ASD), Genes, Gene Selection.

DOI Number: 10.14704/nq.2020.18.11.NQ20228 **NeuroQuantology 2020; 18(11):01-13**

Introduction

Chemical imbalance Spectrum Disorder (ASD) is a mental health issue that outcomes in deferred and irregular improvement. ASD impacts the sensory system and influences the general psychological, passionate, social and physical soundness of the influenced person. The range and seriousness of indications can fluctuate broadly. Normal side effects incorporate trouble with correspondence, trouble with social cooperation's, over the top interests and tedious practices [1].

Corresponding author: Dr.M. Kalaiarasu
Address: [1*]Associate Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore. [2]Associate Professor, Department of Information Technology, Sri Ramakrishna Engineering College, Coimbatore.
[1*]E-mail: kalai.muthusamy@srec.ac.in
[2]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: 04 August 2020 **Accepted:** 30 September 2020

Early acknowledgment, just as social, instructive and family treatments may decrease side effects and bolster improvement and learning. Despite the fact that scientists accept that hereditary variables assume a significant job in the event of the confusion [2-3].

Scholars have endeavored to recognize the most important qualities that can be used as biomarkers for following the confusion. The job of explicit qualities in the improvement of chemical imbalance empowers us to comprehend the system of advancement of the confusion and subsequently anticipate its genuine outcomes [4-5]. Until this point, there is an absence of treatment for the significant manifestations of chemical imbalance, and no precise biomarkers have been recognized in light of the fact that the etiology of mental imbalance isn't plainly known. It has been guaranteed that the total activity of numerous qualities is important to create chemical imbalance issue, a component that adds unpredictability to genomic examinations [6-7].

PC models can be utilized to contemplate mental imbalance using microarray quality articulation information. A microarray is a device that is utilized to gauge whether transformations in explicit qualities are available in a specific person. The most widely recognized kind of microarray is used to quantify quality articulation; right now microarray, the articulation estimations of thousands of qualities are determined from the microarray test [8]. The procedures of AI and information mining are viewed as viable instruments in the utilization of genomic prescription, which utilizes computational strategies and genomic datasets to foresee phenotypes.

For example, Support Vector Machine (SVM) or Gaussian Naïve Bayes (GNB) classifiers have implemented regulated learning strategies in most examinations that combine mind imaging and AI. The subjectivity of highlight determination methodology could be a snag for the correlation of results across contemplates for administered AI techniques. In administered strategies, class marks are doled out to a lot of information utilized as the preparation informational collection; other information focuses (test informational collection) are ordered corresponding to the examples found in the preparation information (utilizing the given names).

At the end of the day, the calculation works to order pre-set up marks (that is, they depend on include choice, or highlight building). The decision of these marks and of the highlights relies upon from the earlier theory or exploratory strategies; thus, they rely upon a degree of subjectivity [9].

As of late, in an open-source Python bundle, valuable computational tools were suggested to identify exhaustive inherent and client characterised highlights for DNA, RNA and protein successions; these are known separately as DNA (repDNA)[10], repRNA[11] and Psein-One[12] portrayals. As for RNA testing, another repRNA was developed to meet growing requirements and to speed up genome research.

Notwithstanding, quality articulation in mental imbalance shows some particular attributes that make quality choice, model creation and expectation more testing than quality articulation investigation of malignant growths. The serious issue in the quality articulation examination of ASD is the trouble in choice and distinguishing proof of the qualities that are generally applicable to chemical imbalance. This issue exists in light of the fact that the quality articulation levels in chemical imbalance issue show significant vacillation among people and on the grounds that the successions of a few of these qualities are exceptionally factor. For quality determination and arrangement in chemical imbalance issues, Geometric Binary Particle Swarm Optimization (GBPSO) with Support Vector Machine (SVM) classifier is used in the current work. It is brought about low precision if the dataset tests are huge. To deal with this issue, Modified Cuckoo Search-Support Vector Machine (MCS-SVM) calculation is proposed to improve the exactness level of classifier.

Literature Review

Zhou et al [13] proposed a chart hypothesis and AI examination of multi-parametric MRI information to improve portrayal and forecast in Autism Spectrum Disorder (ASD). The multi-focus Functional Connect me Project selected details from 127 children with ASD (13.5±6.0 years) and 153 age-and sex-coordinated regularly creating youngsters (14.5±5.7 years). For grouping and forecasting phenotypic highlights that integrated the chemical imbalance symptomatic perception strategy, reconsidered mental imbalance demonstrative meeting, and IQ ratings, an integrative model of 22 quantitative imaging highlights were used.

In 22 imaging highlights, four (caudate-cortical utilitarian availability, caudate volume, and substandard frontal gyrus practical network) were seen as profoundly educational, uniquely improving order and expectation precision when contrasted and the single imaging highlights. This methodology might fill in as a biomarker in visualization, conclusion, and observing infection movement.

Moradi et al [14] proposed a novel methodology for ASD finding and exhibit its handiness with the Autism Brain Imaging Data Exchange (ABIDE) database. Foresee side effect intensity based on estimates of cortical thickness from 156 individuals with ASD from four separate destinations. There are two main stages of the proposed approach: a space adjustment method using Partial Least Squares relapse (PLS) to extend the accuracy of imaging knowledge across local locations; also, A consolidating Support Vector Regression (SVR) learning stage for provincial severity forecast with versatile net punished Linear Regression (LR) for coordinating territorial expectations into an expectation of full seriousness of mind. This demonstration of the effectiveness of the proposed approach to differentiating simple mind anomalies in ASD from the multi-site, multi-convention ABIDE dataset demonstrates the potential to prepare AI strategies to overcome agglomerative knowledge difficulties.

Al Diabat and Al-Shanableh [15] presented another structure for Ensembles Learning ASD screening is considered called Ensemble Classification for Autism Screening (ECAS). ECAS uses an amazing learning methodology that takes into account the creation of various classifiers from verifiable cases and controls and then uses these classifiers to predict medically introverted features on test occasions. The results showed that ECAS had the option of generating better classifiers from the dataset of young people than the other Machine Learning strategies considered with regard to affectability, specificity, and accuracy levels.

Altay and Ulas [16] proposed a characterization strategy for In children aged 4-11 years, ASD determination was utilized. The calculations of Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (K-NN) were used. In order to validate the estimates, 30 per cent of the information index was selected as test data and 70 per cent as information planning. Measurements were calculated by precision, affectability, specificity, accuracy and F-measure estimates, and True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) estimates were used in analysis. LDA calculation reached a notable accuracy while K-NN calculation was lesser accurate.

Thabtah and Peebles [17] presented another Rules-Machine Learning (RML) classifier to improve the exactness and proficiency of ASD location. AI offers propelled systems that develop computerized classifiers that can be abused by clients and clinicians to essentially improve affectability, explicitness, exactness, and proficiency in analytic revelation. Observational outcomes on three datasets identified with youngsters, youths, and grown-ups show that RML offers classifiers with higher prescient precision, affectability, symphonious mean, and explicitness than those of other ML approaches, for example, Boosting, Bagging, choice trees, and rule enlistment.

A decision tree algorithm was proposed to analyze similar factors in datasets collected from the National Autism Research Database (NDAR) consisting of approximately 3000 ASD diagnostic individuals by Hassan and Mokhtar [18]. They were able to classify 15 medical conditions in patients that were highly correlated with ASD diagnoses, applying this study to patients' family medical history, and reporting six ASD-related potentially inherited medical conditions. Literature confirmed their results, thereby opening the way for new directions in the use of decision tree algorithms to understand autism better.

Namara et al [19] proposed to assess the exactness of ASD utilizing two AI grouping calculations are Decision Tree (DT) and Random Forest (RF). The information cleaning and pre-handling steps taken to set up the dataset for grouping, for example, managing missing qualities, exception evacuation, variable determination, and the dividing of the information into preparing and testing subsets. At last, improve expectation precision by using a few distinctive order calculations.

Omar et al [20] proposed a successful forecast model dependent on ML procedure and to build up a portable application for anticipating ASD for individuals of all ages. By combining RF-Iterative Dichotomiser (ID3) and RF-Classification And Regression Trees (CART), a chemical imbalance forecast model was developed and a flexible implementation depending on the proposed expectation model was also created. The

assessment results demonstrated that the proposed expectation model give better outcomes as far as exactness, specificity, affectability, accuracy and False Positive Rate (FPR) for the two sorts of datasets.

Proposed Methodology

For gene subset, level of accuracy is enhanced by proposed Modified Cuckoo Search-Support Vector Machine (MCS-SVM) algorithm. There are three major process involved in this. They are pre-processing, selection of gene and classification. In autism and control classes, gene expression similarity is checked in entire dataset in first step. Removed the genes having unity or close to unity median or mean values. In dataset, gene count is reduced to 9454 from 54,613.

Two parts are formed by dividing the reduced dataset in second step. For training and validating the model, 85% of dataset is used. Another 15% of reduced dataset is used for testing the model for classification of gene. 10-fold run evaluation is used for selecting 200 discriminative features by parallel application of three filters. They are Wilcoxon Rank Sum test (WRS), feature COR relation (COR)and T-Test (TT).

MCS-SVM algorithm is used for selecting most discriminative gene subset in last stage and according to selected gene set, performed the classification. In 10-fold cross validation method, validation and training is performed using this selected gene set in SVM classifier. Experimental setup used for selecting genes related to autism and performing classification is shown in figure 1.

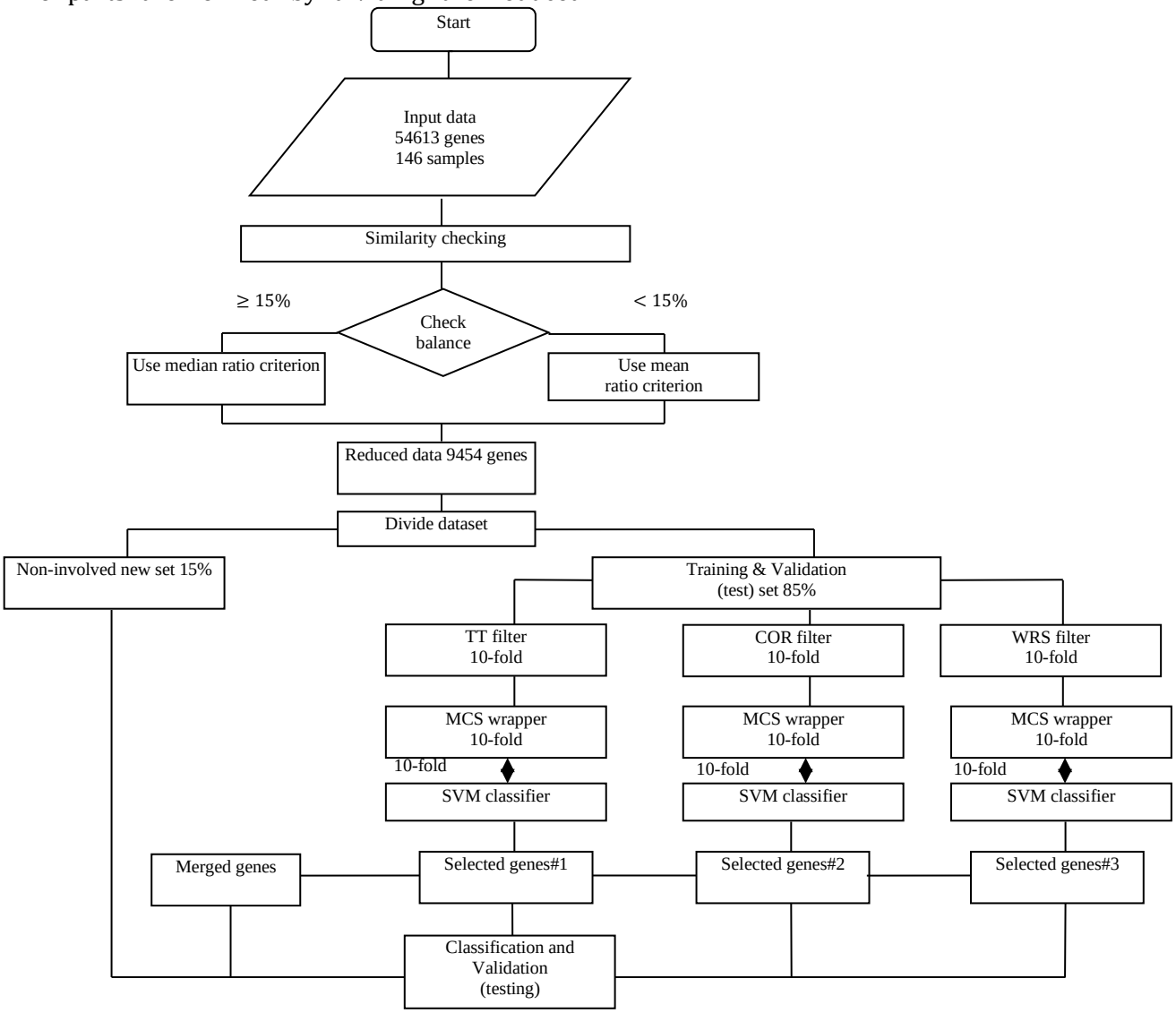

Figure 1. Experimental setup used for selecting genes associated with autism and classifying

Preprocessing Operations

In this study, in the dataset of autism gene expression, most apparent problem is high variance [21]. There 54,613 genes in this dataset of high dimensional. Same non-autism-related and autism-related genes expression is exist only in 146 samples. Genes showing same expression in individuals both sets are useless as per statistical theory. So, in discriminant gens, they will not be included. Subsequent steps are facilitated by removal of this similar genes. This is very useful in selection process of genes in proposes method.

Individual genes expression of median and mean are affected by these similar genes having high variance. Next filter steps may be affected by this. In order to remove same genes, mean ration is used as a basis. When compared to the application of mean criteria, median application is the better choice when feature has outliers as proven [22]. Variance affect the gene expression's mean value.

In every class, identified the genes having very high variance as per its expression by using a new method. Population are not presented reliably by mean values produced because of high variance. In feature selection, undesirable results are produced by this. In various contexts, applied median and mean ratio for avoiding these problems in this study and facilitating the analysis in next step.

Within the class, for gene expressions which do not have high variance, used a mean values ratio and applied a median value ratio for the expressions having high variance. Problems included in dataset's variance are rectified by this method in a creative way. Based on variance, two groups of features are formed by dividing every class features. Separated the low variance and high variance feature set. Threshold value is considered as 15% for division.

By expelling characteristics that have very similar medians in two classes as well as those that have very comparative methods in two classes, e analysis decreases high-dimensional qualities of the dataset. Values in the range of 0.95 and 1/0.95, middle and mean ratios for the two groups are expelled from the dataset. This edge go is deliberately selected to expel non-huge characteristics from the entire dataset to reduce the effect of high-change characteristics.

Selection Using Statistical Filters

The reduced arrangement of characterization in previous developments is used as a contribution to three methods that depend on the channel approach. The factual channels are the two-samples: T-Test (TT), highlight class COR relation (COR) and Wilcoxon Rank Sum test (WRS). The explanation behind picking multi-channel the techniques have diverse relative force, utilization of a blend of these strategies may yield preferable choice execution over the utilization of a solitary channel [22].

Dataset was divided into two sections: one section consisted of 85% of the data and was used for the preparation and approval (testing) of the model; the other section, with 15% was separate as a non-included collection for another genuine world dataset with quality grouping. The best methodology is to apply preparing and 10-overlay approval to abstain from overfitting. Additionally, usage of the entire dataset for include choice delivers a one-sided result that doesn't show the genuine capacity of the model during the test stage. In this manner, the factual separating was rehashed in 10-crease runs on the prepared dataset. In every technique, the places of the separated qualities in every one of the runs were thought about dependent on their position loads.

Next, the weight esteems were added, and the qualities were requested from most ascribed to least credited by the last positions accomplished inside the 10-crease runs. The condition utilized right now a worldwide weight condition that is given by

$$w(f) = \sum_{i=1}^{K} w_i(f)$$

Where every I in K = the quantity of current overlay emphasess in the entire 10-overlap run. The primary applied channel was the t-test, which is a univariate channel includes determination that is frequently utilized in class applications executed in parallel [23]. The regular presumption of the t-test is that the qualities for the two looked at gatherings of qualities are typically appropriated. The invalid theory of the t-test accept equivalent methods and equivalent fluctuations, and the elective speculation dismisses this presumption. The condition of the t-test is

$$t = \frac{c_1 - c_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Where n and m mean the populace sizes of the first and second classes, separately. The consequence of the appraisal calls t, which is equivalent to 1 or 0; 1

speaks to the dismissal of the invalid speculation at the 5% importance level and 0 indicates the acknowledgment of the invalid theory at a similar criticalness level. For the mental imbalance dataset, a typical dissemination of the communicated qualities isn't ensured because of the nearness of exceptions. Accordingly, non-parametric performance of t-test was done in MATLAB by accepting inconsistent differences in the two classes.

The t-test has for some time been utilized in the utilization of microarray highlight choice [23]. It has incredible adaptability when the quantity of highlights is high. The t-test is utilized as a channel followed by wrapper-based quality determination; after this, the grouping calculation is applied. The second applied channel technique was include COR relation with class (COR), a univariate channel highlight choice strategy that can be utilized as a pre-choice advance in microarray quality determination [24]. The estimation of highlight separation, S(f) is communicated by

$$S(f) = \frac{\sum_{k=1}^{K} P_k (c_k - c)^2}{\sigma^2(f) \sum_{k=1}^{K} P_k (1 - P_k)}$$

Where c - mean of an incentive for quality between two classes, c_k - mean for an incentive of the kth class quality, σ^2 (f) - quality change, and P_k - likelihood of appearance of the kth class in the dataset. A high estimation of S(f) speaks to great separation ability of highlight f in recognizing a specific class from other K classes. The third applied channel strategy was Wilcoxon Rank Sum test (WRS). WRS test is a non-parametric channel strategy, it isn't important for the quality articulation information in the classes to be ordinarily conveyed.

Hence, it seems increasingly necessary to apply WRS test to the present dataset. The norm used for this test relies on the middle value to recognize classes. The test analyses example' medians and generates the result as a position instead of a number. The location and calculation of the game plan is resolved by arranging the results of the climbing request. The WRS test considers as the invalid theory the speculation that all qualities start from one class [25]. The measurable recipe of the Wilcoxon rank aggregate is as per the following:

$$s(g) = \sum_{i \in N_0} \sum_{j \in N_1} I((x_j^{(g)} - x_i^{(g)})) \leq 0$$

Where I - function used to distinguish the classes. When the logical expression $(x_j^{(g)} - x_i^{(g)}) \leq 0$ is true, I is 1 else it is 0. $x_i^{(g)}$ - expression value of gene g in sample I, N_0 and N_1-the no of observations in each of the two classes, s(g) - difference in expression of the gene in the two classes and if s(g) becomes 0 or reaches the maximum of $N_0 \times N_1$, the considered gene is ranked in important in classifications. The following equation is used to calculate the gene's importance:

$$q(g) = \max(s(g), N_0 \times N_1 - s(g))$$

Usually, at the end of the analysis, WRS will give the rank of the genes, beginning with the most discriminative genes and proceeding to the less discriminative ones.

Table 1. Similarity Percentages for Sets of 200 Discriminative Genes Selected by Various Filtering Methods

Filtering method	TT	COR	WRS
TT	100	93	69
COR	93	100	69
WRS	69	69	100

Selection Using MCS Algorithm

a) Cuckoo Search

A meta-heuristic calculation rising from the feathered cuckoo is a Cuckoo Quest (CS); winged beings of the "Brood parasites". It never manufactures its own number of qualities and lays their grouping exactness in the quantity of qualities another host fledgling number of qualities. Cuckoo is a most popular brood parasite.

Cuckoo Breeding Behaviour

Cuckoo is intriguing flying creatures, not just in light of the lovely sounds they can make, yet in addition as a result of their forceful multiplication system. Cuckoos lay their grouping exactness in common number of qualities, however they may evacuate others arrangement precision to expand the bring forth likelihood of their own characterization precision. Many animal groups attract parasitism from the dedicated brood by setting their arrangement accuracy in the amount of characteristics of other host winged creatures (regularly distinct species) [26]. There are three main forms of parasitism of the brood: intra-specific parasitism, helpful reproduction, and takeover number. Some host flying creatures link the barging in cuckoos to a direct clash. In the event that a host fowl finds the eggs are not their claims, they will either dispose of these outsider arrangement precision or basically relinquish its

home and fabricate another number of qualities somewhere else. This decreases the likelihood of their grouping exactness being surrendered and along these lines builds their reproductively. Likewise, the planning of characterization precision laying of certain species is additionally stunning.

Parasitic cuckoos regularly pick various qualities where the host winged animal simply laid its own arrangement precision. All in all, the cuckoo order precision bring forth marginally sooner than their host characterization exactness. When the main cuckoo chick is incubated, the principal intuition move it will make is to remove the host grouping, which expands the cuckoo a lot of nourishment gave by its host winged animal. Concentrates likewise show that a cuckoo chick can likewise impersonate the call of host chicks to access all the more bolstering open door [27]. L'evy Flights Different investigations have demonstrated that the flight conduct of numerous creatures and creepy crawlies may represent some commonplace attributes of L'evy flights. Concentrates on human conduct such rummaging examples likewise show the commonplace component of L'evy flights. Indeed, even light can be identified with L'evy flights. In this manner, such conduct has been applied to enhancement and ideal pursuit, and starter results show its promising capacity.

Cuckoo Search

In various qualities, every arrangement precision speaks to an answer and cuckoo grouping exactness speaks to another and great arrangement. The acquired arrangement is another arrangement dependent on the current one and the alteration of certain qualities. In the least complex structure each number of qualities has one arrangement exactness of cuckoo in which each number of qualities will have numerous characterization precision speaks to a lot of arrangements. CS is effectively used to tackle booking issues and used to take care of plan advancement issues in basic building. Cuckoo search glorified such rearing conduct and can be applied to different advancement issues [28].

Each cuckoo lays each arrangement precision in turn and dumps it in a haphazardly picked number of qualities.

The best number of qualities with the high caliber of arrangement precision will convey to the following ages.

The quantity of accessible host number of qualities is fixed and if a host winged animal recognizes the cuckoo order precision with the likelihood of p_a=[0,1] then the host fledgling can either discard them or desert them and assemble another chose qualities.

As a further guess, this last presumption can be approximated by a portion p_a of the n have number of qualities are supplanted by new number of qualities (with new irregular arrangements). For an augmentation issue, the quality or wellness of an answer can essentially be relative to the estimation of the goal work. Different types of wellness can be characterized along these lines to the wellness work in hereditary calculations.

For the usage perspective, arrangement precision in various qualities speaks to an answer, and each cuckoo can lay just one egg (along these lines speaking to one arrangement), the point is to utilize the new and possibly better arrangements (cuckoos) to supplant a not all that great arrangement in the quantity of qualities. Clearly, this calculation can be reached out to the more convoluted situation where each number of qualities has different order exactness speaking to a lot of arrangements, or speaking to multi-destinations [28]. In view of these three standards, the essential strides of the Cuckoo Search (CS) can be condensed as the pseudo code.

Cuckoo Search via Levy Flights

Objective function $f(x), x = (x_1, \ldots\ldots, x_d)^T$

Generate initial population of n host number of genes x_i

While $(t < MaxGeneration)$ or (stop criterion)

Get a cuckoo randomly/generate a solution by Levy flights

and then evaluate its quality/fitness F_i

Choose a number of genes among $n(say, j)$ randomly

if $(F_i > F_j)$

Replace j by the new solution

end

Abandon a fraction (p_a) of worse number of genes& generate new solutions

Keep selected genes solutions and find the current number of genes

end while

Post process results and visualization

This algorithm uses a balanced combination of a local random walk and the global explorative random walk, controlled by a switching parameter p_a.

The local random walk can be written as

$$x_i^{t+1} = x_i^t + s \otimes H(p_a - \varepsilon) \otimes (x_j^t - x_k^t)$$

where x_j^t and x_k^t are two different solutions selected randomly by random permutation, $H(u)$ is a Heaviside function, ε is a random number drawn from a uniform distribution, and s is the step size. On the other hand, the global random walk is carried out by using L´evy flights

$$x_i^{t+1} = x_i^t + \alpha L(s, \lambda)$$

Where

$$L(s, \lambda) = \frac{\lambda \Gamma(\lambda) \sin\left(\frac{\pi\lambda}{2}\right)}{\pi} \frac{1}{S^{1+\lambda}}, (s \gg s_0 > 0)$$

Random walk is provided by L´evy flight. From L´evy distribution, step length of this random walk is computed,

$$Levy \sim \frac{1}{S^{\lambda+1}}, (0 < \lambda \leq 2)$$

Here the means basically structure an irregular walk process with a force law step-length circulation with a substantial tail. A portion of the new arrangements ought to be created by L'evy stroll around the best arrangement acquired up until now, this will accelerate the nearby hunt. Nonetheless, a considerable part of the new arrangements ought to be produced by a wide margin field randomization and whose areas ought to be far enough from the present best arrangement, this will ensure that the framework won't be caught in a neighborhood ideal.

Drawbacks of Cuckoo Search:

(i) Boundary Issue: The areas of some number of qualities might be out of the limit; when this happens CS calculation utilizes the limit an incentive to supplant this area. The bound managing technique will bring about a great deal of number of qualities at a similar area on the limit, which is wasteful.

(ii) CS strategy isn't great; it effectively falls into the neighborhood ideal arrangement and the moderate pace of intermingling.

To unravel these issues proposed a Modified Cuckoo Search (MCS) calculation.

b) Modified Cuckoo Search (MCS)

In reality, if a cuckoo's order exactness is fundamentally the same as a host's grouping precision, at that point this current cuckoo's arrangement precision is more averse to be found, subsequently the wellness ought to be identified with the distinction in arrangements. Consequently, it is a smart thought to do an arbitrary stroll in a one-sided path with some irregular advance sizes. Both, unique, and changed code utilize arbitrary advance sizes. In the first code, step size is determined utilizing following code articulation:

$$r * number\ of\ genes[permute1[i]][j] - number\ of\ genes[permute2[i]][j]$$

Where, random number is represented as r and it lies between [0,1] range, number of genes is matrix which contains selected genes along with their variables, permute1 and permute2 are different rows permutation functions applied on number of genes matrix.

$$r * number\ of\ genes[sorted[i]][j] - number\ of\ genes[permute[i]][j]$$

The thing that matters is that rather than permute1, utilized arranged capacity. This capacity sorts number of qualities network by wellness of contained arrangements. Right now, wellness arrangements have slight bit of leeway over arrangements with lower wellness. This strategy keeps the choice weight (how much profoundly fit arrangements are chosen) towards better arrangements and calculation ought to accomplish better outcomes [29]. That doesn't imply that high wellness arrangements will flood populace and the calculation will get stuck in neighborhood ideal. CS calculation distinguishes best arrangement Xbest toward the start of every iterative advance.

$$\sigma_u = \left\{ \frac{\Gamma(1+\beta) \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left[\frac{(1+\beta)}{2}\right] \beta 2^{\frac{(\beta-1)}{2}}} \right\}, \sigma_v = 1$$

Where, Lévy distribution parameter is represented as β and gamma function is represented by Γ. With donor vector v, cuckoo I's evolution starts, where, $v = x_i^{(t)}$. Following expression is used for computing step size,

$$Stepsize^{(t+1)} = 0.01 \frac{u^{(t+1)}}{\left|v^{(t+1)}\right|^{\frac{1}{\beta}}} (v - x_{best})$$

Where $u \sim N(0, \sigma_u^2)$ $v \sim N(0, \sigma_v^2)$ are samples from corresponding normal distributions. For the progression size as per Levi appropriation utilized suggested parameter β=1.5. The best discovered arrangement however it isn't recalled in any different memory since the best arrangement, as indicated by the calculation, is constantly held among current arrangements. That implies that the procedure holds Markov property since the following state again relies just upon the present state and not the past. Nonetheless, the

randomization is progressively productive as the progression length is overwhelming followed, and any enormous advance is conceivable.

Classification Using Support Vector Machine (SVM)

MCS is used as a wrapper that contains a Support Vector Machine (SVM) determination technique. The SVM calculation is used in view of the fact that, despite the accessibility of minimal planning tests, it can provide high-dimensional details with fair order accuracy. SVM are a gathering of directed AI strategies known as a help vector arrange which are applied in assortment of quality arrangement of illnesses.

In addition, SVM can perform both direct and nonlinear grouping of separate information. In a straightforward case, the direct option limit is executed to the degree that the smallest separation between the planning tests and the limit (edge) is increased. The preparation data tests are known as assist vectors near the class limit and along the hyperplanes. After mapping the quality space of low dimensionality separated from the info space into a quality space of high dimensionality to achieve productive order, SVM can deal with nonlinear information [30].

Another property of SVM is that the quantity of coefficients to be resolved is basically subject to the quantity of tests as opposed to on the quantity of qualities. This is a valuable quality of SVM for microarray information because of the nearness of a low proportion of tests to qualities right now dataset. In any case, it has been demonstrated that diminishing the quantity of qualities expands SVM execution. To use SVM as a characterization calculation in quality articulation and succession datasets, part works are typically utilized. This permits the client to get a symmetrical hyperplane to recognize the qualities in a particular measurement.

SVM classifier is trained using a Radial Basis Function (RBF) kernel function in this study. For autism dataset, due to high accuracy in classification, polynomial kernel is used. Kernel parameters are optimized further. Amino acid distance pairs and alphabet profile which are reduced are incorporated with general pseudo amino acid for DNA-binding proteins identification. From frequency profiles, extracted a evolutionary information by sequence-based kernels combination which are performing well. They are SVM-LA, SVM-pair wise and SV-Ngram.

Due to high accuracy in classification, polynomial kernel is used in SVM in this study. Feature subset is selected by using fitness function of MCS. Accuracy of SVM classifier computed so far is used in this process. Better accuracy and best gene subset is defined by this method. In 10-fold cross validation, this is applied. In best gene computation process, used this entire training set.

Results and Discussion

Autisim microarray dataset is used for experimental analysis and comparison. From, public repository GEO (NCBI), there are downloaded. There are 46 samples and 54,613 features in this dataset. There are two classes in dataset samples. 69 samples are control class, 77 samples are autisim class. In Phoenix area, from persons, these samples are collected form control and autistic individuals [31].

LR-SVM is used for comparing performance of proposed MCS-SVM. On a test dataset, classification model performance is described by using a table called confusion matrix. True values are computed using this. Algorithm performance can be visualized using this. Table 2 shows confusion matrix. It consist of two row and columns, They represent true negatives, true positives, false negatives and false positives.

Table 2. Confusion Matrix

		Predicted	
		Positive	**Negative**
Actual	**Positive**	tp	fp
	Negative	fn	tn

- True Positive (tp)- correctly predicted label,
- True Negative (tn) - correctly predicted other label
- False Positive (fp) - falsely Predicted label
- False Negative (fn) - missing and incoming label.

Ratio between expected positive cases to correctly predicted cases defines Precision (P). It is given by,

$$\text{Precision (P)} = tp/(tp+fp) \quad (1)$$

Ratio between positive cases to total number of cases defines True Positive Rate (TPR) or recall. It is given by,

$$\text{TPR} = tp/(tp+fn) \quad (2)$$

Harmonic mean of recall and precision is termed as balanced F-score and F-measure or F1 score. Constant value 2 is multiplied with this to make a score F1 score of 1 while precision and recall takes a value of 1. It is expressed as,

F-score= 2tp/(2tp+fp+fn) (3)

Ratio between total corrections to total number of predictions defines accuracy value. It is represented in expression (4),

Accuracy= (tp+tn)/ (tp+tn+fp+fn) (4)

Table 3 list the comparison of performance of classification methods.

Table 3. Performance Comparison Metrics vs. ASD Classification Techniques

Methods	Metrics			
	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
LR	71.4	76.4	73.8	68.5
GBPSO-SVM	81.4	77.8	79.5	75.3
MCS-SVM	85.3	89.7	87.5	86.4

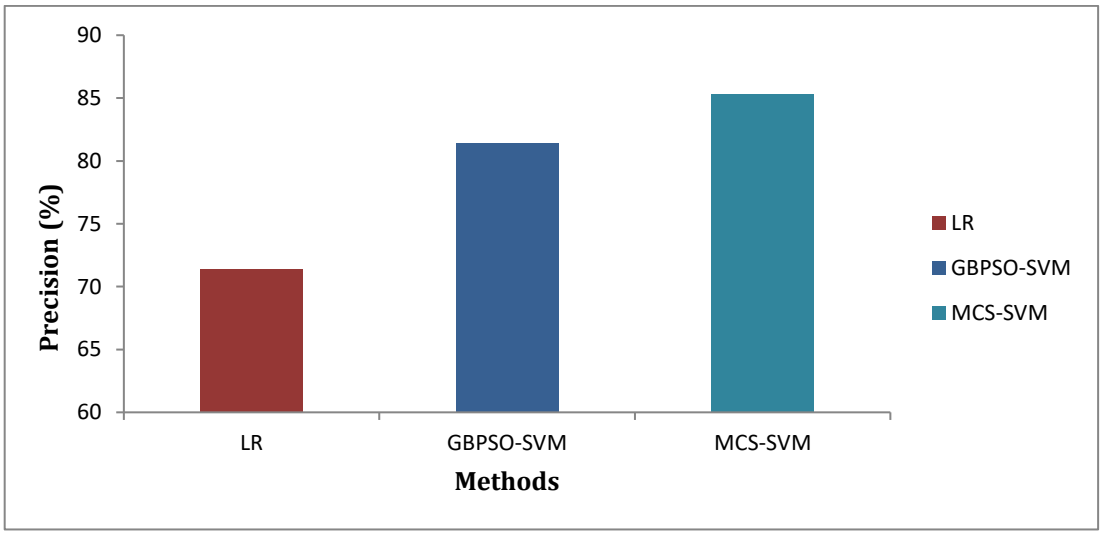

Figure 2. Precision Results Evaluation of ASD Classification Methods

The comparison of precise value results of MCS-SVM, GBPSO-SVM and SV-SVM classification techniques is shown in Figure 2. As shown by the results, about 85.3% of accuracy outcomes are provided by the proposed MCS-SVM. Where, GBPSO-SVM produces 81.4% accuracy and 71.4% of accurate results come from LR-SVM methods.

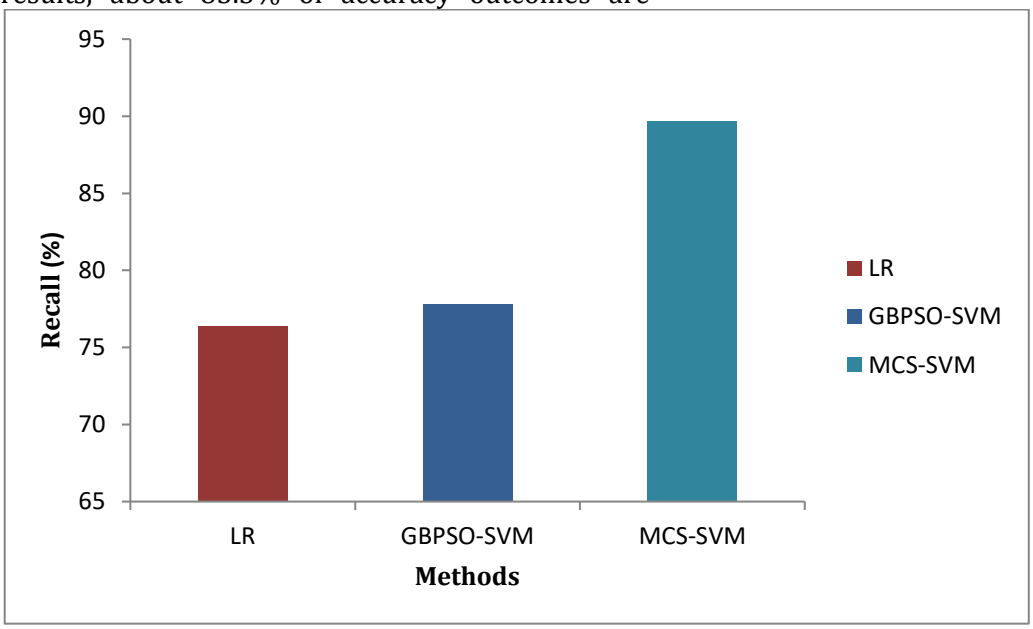

Figure 3. Recall Results Evaluation of ASD Classification Methods

The comparison of recall values of MCS-SVM, GBPSO-SVM and SV-SVM classification techniques is shown in Figure 3. The suggested MCS-SVM produces 89.7 % of the recall results, Where, GBPSO-SVM scores 77.8 % on recall values and LR-SVM scores 76.4%.

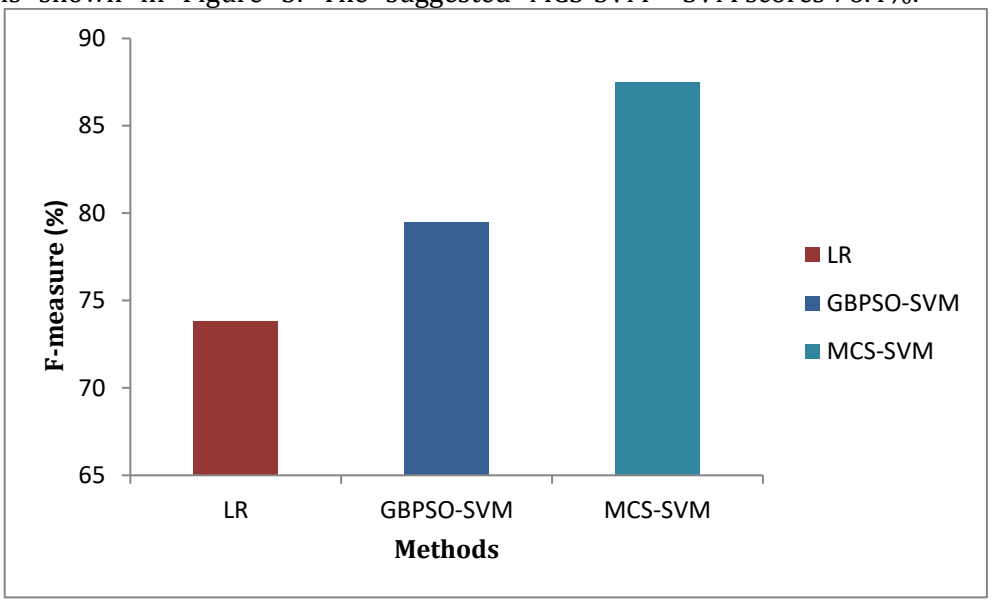

Figure 4. F-measure Results Evaluation of ASD Classification Methods

Figure 4 shows the comparison of results of F-measure value of MCS-SVM, GBPSO-SVM and SV-SVM classification techniques. Around 87.5% of f-measure results are produced by proposed MCS-SVM as shown by results. Where, 795% of f-measure results are produced by GBPSO-SVM and 73.8%f-measure results are produced by LR-SVM methods.

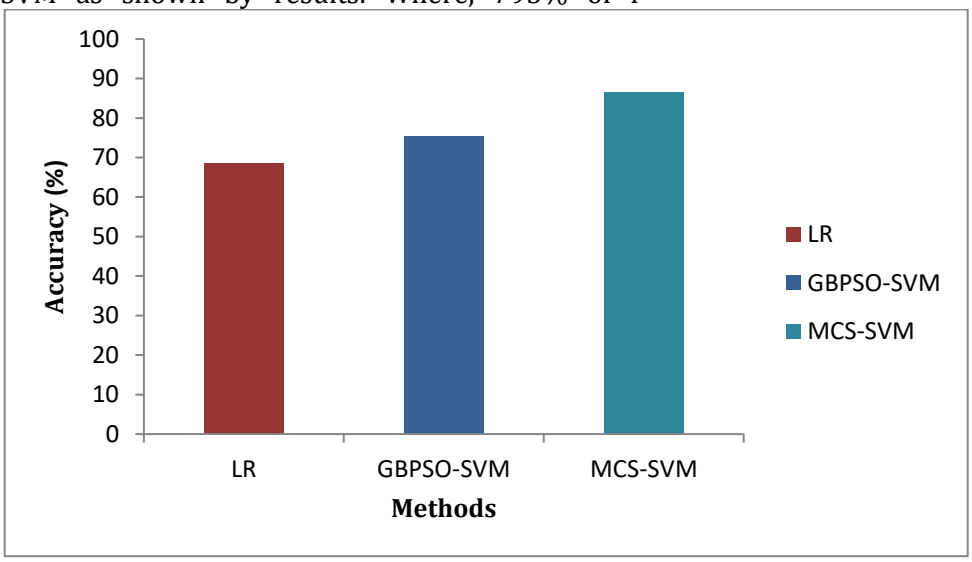

Figure 5. Accuracy Results Evaluation of ASDClassification Methods

Figure 5 shows the comparison of results of accuracy value of three classification techniques. Around 86.4% of accurate results are produced by proposed MCS-SVM as shown by results. Where, 75.3% accurate results are produced by GBPSO-SVM and 68.5% accurate results are produced by LR-SVM methods.

Conclusion and Future Work

Autism Spectrum Disorder (ASD) is an extremely important neurological disorder that disrupts the social communication skills of an individual. It is useful to develop new screening approaches using structural and functional brain networks as it can be helpful for the patient to diagnose autism with plausible specificity at a young age, though signs

are not yet evident.

ASD's consistency articulation figures were essentially broken down with the goal of enhancing the method of selection and order. A combination of factual channels and a wrapper-based MCS-SVM calculation was used to cultivate this. It was noted that the declaration of characteristics likely associated with ASD varies exceptionally amongst perceptions; henceforth, the usage of the mean proportion foundation alone to evacuate comparable qualities doesn't give a dependable outcome. For gene subset, level of accuracy is enhanced by proposed Modified Cuckoo Search-Support Vector Machine (MCS-SVM) algorithm. There are three major process involved in this. They are pre-processing, selection of gene and classification. In autism and control classes, gene expression similarity is checked in entire dataset in first step. Removed the genes having unity or close to unity median or mean values. In dataset, gene count is reduced to 9454 from 54,613.

Two parts are formed by dividing the reduced dataset in second step. For training and validating the model, 85% of dataset is used. Another 15% of reduced dataset is used for testing the model for classification of gene. 10-fold run evaluation is used for selecting 200 discriminative features by parallel application of three filters. They are Wilcox on Rank Sum test (WRS), feature Correlation (COR) and T-Test (TT).

The MCS-SVM algorithm is used to pick the most discriminatory gene subset in the last stage and classification is carried out according to the selected gene collection. Validation and training was performed using this selected gene set in the SVM classifier in a 10-fold cross validation process. An autism microarray dataset that was downloaded from a public repository GEO (NCBI) comprised of experimental data used in the study. The proposed MCS-SVM classifier provides higher accuracy value compare with other existing methods such as LR, GBPSO-SVM.

In future work ASD research endeavors.

(i) The other analytical methods such as prescriptive a, descriptive, predictive methods can be applied on the data.

(ii) In autism, within a formal diagnostic tool, ensemble learners can be embedded.

(iii) In order to classify autistic children, high accuracy classifiers are combined to build a model. In autistic children, autism level can be predicted using this model.

(iv) An efficient way of detecting the autism using a combined approach of two segmentation techniques. So that the combination of two segmentation method will help to produce a new efficient and accurate algorithm in order to the detection.

References

De Rubeis S, Buxbaum JD. Recent advances in the genetics of autism spectrum disorder. *Current neurology and neuroscience reports* 2015; 15(6): 36.

Thurm A, Swedo SE. The importance of autism research. *Dialogues in clinical neuroscience* 2012; 14(3): 219-222.

Yoo H. Genetics of autism spectrum disorder: current status and possible clinical applications. *Experimental neurobiology* 2015; 24(4): 257-272.

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.

Munoz R, Morales C, Villarroel R, Quezada Á, De Albuquerque VHC. Developing a software that supports the improvement of the theory of mind in children with autism spectrum disorder. *IEEE Access* 2018; 7: 7948-7956.

Javed H, Jeon M, Howard A, Park CH. Robot-assisted socio-emotional intervention framework for children with Autism Spectrum disorder. *In Companion of the ACM/IEEE International Conference on Human-Robot Interaction* 2018: 131-132.

Brahim A, El Hassani MH, Farrugia N. Classification of autism spectrum disorder through the graph fourier transform of fMRI temporal signals projected on structural connectome. *In International Conference on Computer Analysis of Images and Patterns*, *Springer, Cham* 2019: 45-55.

Wei W, Liu Z, Huang L, Nebout A, Le Meur O. Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder. *In IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* 2019: 621-624.

Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuro Image: Clinical* 2018; 17: 16-23.

Liu B, Liu F, Fang L, Wang X, Chou KC. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* 2015; 31(8): 1307-1309.

Liu B, Liu F, Fang L, Wang X, Chou KC. repRNA: a web server for generating various feature vectors of RNA sequences. *Molecular Genetics and Genomics* 2016; 291(1): 473-481.

Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic acids research* 2015; 43(W1): W65-W71.

Zhou Y, Yu F, Duong T. Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning. *PloS one* 2014; 9(6): e90405.

Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *Neuroimage* 2017; 144: 128-141.

Al Diabat M, Al-Shanableh, N. Ensemble Learning Model for Screening Autismin Children. *International Journal of Computer Science &Information Technology (IJCSIT)* 2019; 11(2): 45-62.

Altay O, Ulas M. Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. *In 6th International Symposium on Digital Forensic and Security (ISDFS)* 2018: 1-4.

Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. *Health informatics journal* 2019.

Hassan MM, Mokhtar HM. Investigating autism etiology and heterogeneity by decision tree algorithm. *Informatics in Medicine Unlocked* 2019; 16: 100215.

McNamara B, Lora C, Yang D, Flores F, Daly P. Machine Learning Classification of Adults with Autism Spectrum Disorder 2018. http://rstudio-pubs-static.s3.amazonaws.com/383049_1faa93345b324da6a1081506f371a8dd.html

Omar KS, Mondal P, Khan NS, Rizvi MRK, Islam MN. A machine learning approach to predict autism spectrum disorder. *In International Conference on Electrical, Computer and Communication Engineering (ECCE)* 2019: 1-6.

El-Fishawy P. The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatric Clinics* 2010; 33(1): 83-105.

Latkowski T, Osowski S. Computerized system for recognition of autism on the basis of gene expression microarray data. *Computers in biology and medicine* 2015; 56: 82-88.

Huertas C, Juárez-Ramírez R. Filter feature selection performance comparison in high-dimensional data: A theoretical and empirical analysis of most popular algorithms. *In 17th International Conference on Information Fusion (FUSION)* 2014: 1-8.

Muszyński M, Osowski S. Data mining methods for gene selection on the basis of gene expression arrays. *International Journal of Applied Mathematics and Computer Science* 2014; 24(3): 657-668.

Khoshgoftaar T, Dittman D, Wald R, Fazelpour A. First order statistics based feature selection: A diverse and powerful family of feature seleciton techniques. *In 11th International Conference on Machine Learning and Applications* 2012; 2: 151-157.

Cheng Z, Wang J, Zhang M, Song H, Chang T, Bi Y, Sun K. Improvement and application of adaptive hybrid cuckoo search algorithm. *IEEE Access* 2019; 7: 145489-145515.

Zhang M, He DX, Zhu CL. Cuckoo Search Algorithm Based on Hybrid-Mutation. *In 12th International Conference on Computational Intelligence and Security (CIS)* 2016: 538-542.

Majumdar D, Mallick S. Cuckoo search algorithm for constraint satisfaction and optimization. *In Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* 2016: 235-240.

Liu L, Liu X, Wang N, Zou P. Modified cuckoo search algorithm with variational parameters and logistic map. *Algorithms* 2018; 11(3): 30-40.

Hassanien AE, Al-Shammari ET, Ghali NI. Computational intelligence techniques in bioinformatics. *Computational biology and chemistry* 2013; 47: 37-47.

Autistic children and their father's age: peripheral blood lymphocytes [Internet]. www.ncbi.nlm.nih.gov 2011. http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431.

Bibliography

Dr.M. Kalaiarasu ME. Ph.D, Working as Associate Professor in Department of Information Technology Sri Ramakrishna Engineering College, Coimbatore – 641022. He has completed his Bachelors of Engineering (computer science Engineering) and his Masters in Engineering (Software Engineering) in Sri Ramakrishna Engineering College, Coimbatore. He has completed his Ph.D (Data Mining) in the year 2015 at Anna University, Chennai. At Present his working areas are Data Mining, Data Analytics, Network Security, Software Engineering.

Dr.J. Anitha is presently working as Associate Professor in the Department of Information Technology at Sri Ramakrishna Engineering College, Coimbatore. She received her Ph.D in the faculty of Information and Communication Engineering from Anna University, Chennai in 2016. Her research interests include Data Privacy, Artificial Intelligence and Natural Language Processing.