Investigating machine learning techniques for the detection of autism using DNA copy number

Fuad M. Alkoot

Higher Institute of Telecommunication & Navigation, PAAET,
Email: fm.alkoot@paaet.edu.kw

Abdullah K. Alqallaf

Electrical Engineering Department, College of Engineering and Petroleum, Kuwait University,
Email: al.qallaf@ku.edu.kw

Abstract:- Automated and speedy autism detection is needed to facilitate urgently required therapy. However, Contrary to cancer, autism detection using microarray genetic data has not attracted much attention. In this paper, we investigate autism detection using machine learning techniques. Here, we study five chromosomal regions associated with behavioral abnormalities. The main goal is to test whether DNA copy number data with machine learning tools can result in an abbreviated and accurate instrument for classification of autism. For that propose a system comprising of four stages is proposed. Where at each stage we experiment with different feature reduction, classification and combination methods to find if it is possible to detect autism using genetic data and to find the methods that yield best detection rates.

The experimental results show that our classifier-based system can achieve optimum accuracy of early screening of the targeted disease. Therefore, through the application of machine learning tools we were able to construct a classifier system that finds if a person will be or is suffering from autism even before any behavioral signs start to appear. We achieved optimum accuracy when tested on independent and unseen test data. The optimum performance of 100% was achieved using our proposed clustering with deleted redundancies feature selection method. The optimum performance was mostly achieved using a three layer neural network back-propagation classifier combined using the feature selection based combiner. The feature size that yields 100% classification rate depends on the chromosome data and the cross-validation iteration. However, for the different chromosomes it ranged between 150 and 500.

Keywords: classifier, autism detection, combining, feature selection, neural network, nearest neighbor, CGH data, DNA copy number variation.

1. Introduction:

Autism spectrum disorder is a life-long brain disorder that is normally diagnosed in early childhood [1, 2]. People with autism have difficulties communicating, forming relationships with others and find it hard to make sense of the world around them. It is varying in severity and impact from individual to individual, ranging from those with no speech and severe learning disabilities to people with IQs in the average range who are able to hold down a job or start a family. People with autism spectrum disorder (ASD) demonstrate significantly challenging behaviors; most need specialist support and care. First identified more than 50 years ago, autism has received a great deal of attention in recent years and it is one of the most common neurological developmental disorders. No one knows exactly why but the brain develops differently in people with autism. The absence of a clear understanding of what causes autism makes finding effective therapies very difficult. It is now widely accepted by scientists that a predisposition to autism is inherited with the underlying genetic cause of up to 40% of autism cases identified up until now [3]. Identification of the condition is at present based solely on observed behaviors. The behavioral method of identification requires experts at special medical centers. The behavioral exam is divided in to four modules where each is geared towards a specific group based on language and age group [4]. This observational test must be run by a certified professional. The time

used from observation to scoring is between 60 to 90 minutes [5]. Due to these limiting requirements families with potentially autistic children that require a diagnosis may wait as long as 13 months [6]. The delay in the detection and diagnosis leads to a delay in the delivery of critically needed speech and behavioral therapies that have significant positive impact on a child's development. Therefore, shorter approaches for early detection are needed.

Genetically, the advancement in technology lead to microarray sequencing which produces high resolution genetic data and lead to the understanding of the complex dynamic interactions between complex diseases and the biological system components of genes and gene products. Genetic variations are found in both apparently normal-species forming their unique features and diseased-species as genetic disorders. These variations may be de-novo and may contribute significantly to disease susceptibility. Further advances lead to a new technique that measures DNA copy number variations (CNVs), where the intensity values represent number of DNA copies at specific genetic positions along the genome. Using machine learning terminology the produced values represent the features in each sample of our data. Similar to most gene based data, the array comparative genomic hybridization (aCGH) [7-9] is a molecular cytogenetic-based approach used to generate the DNA copy number data with a very high dimensionality feature vector that contains much irrelevant information. Therefore, there is a need for an advanced and unique classification system that can use DNA data to identify autism. The system will enable us to improve early screening and diagnosis, in order to achieve timely and effective intervention. However, the data has a very small sample size, leading to the curse of dimensionality. The data is also difficult due to the overlap between class distributions. Therefore, we need to initially preprocess it then use proper feature selection methods to reduce its dimensionality, before any classification system is designed. The need for machine learning tools that deal with the different stages of this problem is clear and obvious.

Machine learning tools are increasingly being used in many application areas to automate decisions. Researchers faced with the task of classification use classifiers or mathematical models that are able to perform the task of classification or decision making, based on a previously provided data. These classifier models or experts have an ability to spot trends and relationships in large data sets which makes them well suited for many applications. In the field of medicine, classifiers are used to accurately classify diseases, genes, tumors, and other medical phenomena [10, 11, 12, 13, 14]. Many have evaluated the performance of classification and feature selection pairs in microarray experiments on cancer detection problems to find the most appropriate machine learning tools. However, the main focus of this study is to use genetic dataset in the form of DNA copy number to detect autism using machine learning techniques. In this type of data the selected features are the genomic position which could include a gene or part of a gene. Details on the dataset, its background and generation are available at [15].

Furthermore, there have been rare attempts to detect autism using machine learning tools. The work by Wall et.al. [16] uses machine learning techniques on behavioral data for autism detection. They experiment with several variations of the decision tree and the nearest neighbor classifiers. Another technique for autism detection is an advanced sensor system that reads brain signals to detect autism, which is still experimental with small success. Stephen Scherers' team attempted to use the genome with the traditional genetic analysis method to distinguish between children with autism, but without success until they found a link between copy number variation and autism [17].

We will experiment with different feature reduction, classification and combination methods to find the best system for autism detection using DNA copy number, GCH, data. The paper is organized as follows. Section 2 presents a background on previous research involving the use of machine learning techniques for detecting diseases using microarray gene expression data. Section 3 our proposed method for detecting autism using gene expression data is presented followed by Section 4 that demonstrates the genomic data description and its preprocessing method. Section 5 presents the experimental

methodology used to extract the biological information and design the system. Section 6 presents the experimental results. The paper is brought to conclusion in section 7.

2. Prior works:

The main focus of this study is to present the machine learning techniques as tools for identifying the association between the targeted disease, autism spectrum disorder (ASD) and the genetic datasets in the form of DNA copy number produced using the CGH method. There is little work on aCGH data in the form of DNA copy number for disease diagnoses using machine learning techniques. Most machine learning techniques were applied to microarray gene expression data. To highlight the novelty of our proposed machine learning technique on high dimensional genetic data we survey prior work on machine learning techniques to identify diseases and disorders using any genetic data including microarray gene expression data.

The literature on cancer detection using genetic microarray data and machine learning tools involve experiments on several research lines. Most worked on finding the best feature reduction, best classification and best combination methods. Others investigated cluster analysis and microarray data integration. Our work involves experiments on finding the feature reduction method suitable for our data, and the classification and combination methods that yield optimum detection rates, given the microarray data specifically prepared for autism detection.

An overview of existing methods of microarray based classifiers is presented by Boulesteix et.al. [18]. Emphasizing the suboptimal procedures in existing methods of classifier evaluation and validation, they address accuracy measures, error rate estimation procedures, variable selection, choice of classifiers and validation strategy. They also address a common mistake by researchers that use the test set with the training set in designing the feature reduction algorithm. They found that many fail to use an untouched test set for predicting the performance of their designed system. They also point out the need to use a 10 fold cross validation. All these recommendations are implemented in our experimental methodology.

The majority of work on using machine learning tools for microarray gene expression data problems focus on proving the advantage of combining when dealing with such problems. As can be gleaned from the literature in Table 1 various machine learning tools and methods have been presented and proposed to detect diseases from microarray gene expression datasets. Most are comparing the performances of different types of feature selection methods, classifiers and/or combiners. They show that some methods are not always successful and suffer from drawbacks. The overwhelming majority of work is for the detection of cancer. No previous work was found for the detection of autism using aCGH data in the form of DNA copy number. However, recently authors in [19, 20] use microarray gene expression data for the recognition of autism. They investigate gene selection methods to find the most representative input attributes for an ensemble of classifiers. They find the contents of small set of the most important genes associated with autism. The selected genes are used in a classifier system to recognize autism. They experiment with SVM classifiers only. In contrast our study is based on using finely-tiled oligonucleotide aCGH microarray data that investigates different feature selection, classifier and combiner methods. Other papers have investigated the integration of microarray data, which is a form of data fusion. Authors in [21, 22, 23] investigated the integration of microarray data to yield improved clustering results. Anna et.al [24] proposed using a formal concept analysis approach for analysis of clustering solutions. They integrate data from different microarray data sets and use clustering to pool together clustering solutions that are further analyzed by the Formal Concept Analysis method, FCA. Yves et.al. [25] overview many clustering methods where some are specific to microarray data then present a clustering algorithm called adaptive quality based clustering that addresses several shortcomings of existing methods. They present a web tool that allows easy analysis of microarray gene data for motif finding. Motifs are overrepresented patterns in DNA.

Table 1 Summary of previous work using machine learning techniques for cancer detection using microarray gene expression data.

Author	Feature selection	classifiers	combining	Data sets	Remarks
Kim et. Al. [26]	7 feature selection methods	MLP, SVM and k-nearest neighbor classifiers	find the best ensemble using an evolutionary algorithm	two cancer data sets, lymphoma and colon	They mix different feature selection methods with different classifiers to make a diverse ensemble
Cho and Won [27]	7 feature selection methods	4 classifiers	They combine the classifiers using majority voting, weighted voting, and Bayesian approach	three cancer: leukemia, colon and lymphoma data set	paired a neural network classifier with principle component analysis feature reduction method for a tumor data. also paired the support vector machine, (SVM), with the signal to noise ratio feature selection method for the ovarian tissue dataset
Cho and Ryu [28]	Feature selection with non overlapping correlation	6 classifier types	Combine pairs of classifiers trained on different feature subsets. voting, weighted voting and a baysian combination	three cancer	
Dudoit et. al. [29]	Not discussed	3 classifiers: LDA, nearest neighbor and decision tree.	Bagging and boosting	leukemia, lymphoma, 60 cancer cell line	
Lee et. al. [30]	3 feature selection methods	21 classification methods		7 cancer data	show the choice of feature selection technique has a large impact on classifier performance
Tan and Gilbert [31]	None	C4.5 decision trees	bagging and boosting	7 cancer	
Dettling [32]	Incorporated in combining	SVM	Bagaboost, RSM, bagging	6 cancers; Leukemia, colon, prostate, lymphoma, tumor, brain	merge bagging with boosting
Dettling and Buhlmann [33]	Feature preselection method; nonparametric rank based equivalent to Wilcoxon's two sample test.	Logitboost, nearest neighbor, decision tree	modify boosting	Leukemia, colon, prostate, lymphoma, tumor, brain	
Valentini et. al. [34]	None	SVM	bagging	leukemia and colon	
Xu and Zhang [35]	Features are selected from dynamically adjusted bootstraps of the training dataset	SVM	bagging	leukemia and colon	Suggest boost feature subset selection
Golub et.al. [36]	Unsupervised class discovery	class discovery procedure to classify		two types of unsupervised leukemia, Lymphoblastic and myeloid	
Loris et.al. [37]	4 different feature selection methods to produce feature subsets for combining	SVM	Feature subsets	several cancer	
Yuehui Chen et.al. [38]	extract genes and reduce dimensionality using a correlation analysis technique	EDA classifiers	Claim a novel ensemble method	4 cancer	
Sung-bae Cho and J. Ryu [39]	several feature selection approaches that are based on correlation analysis or the signal to noise feature selection method	MLP Neural, SVM and kNN	Neural fusion	Lymphoblastic leukemia, myeloid leukemia	present a classification framework that combines a pair of classifiers trained with mutually exclusive features
Sung-Bae Cho and Hong-Hee Won [40]	2 feature selection methods	4 classifiers:, SVM, kNN, MLP and self-organizing maps	Vote, weighted vote and baysian	3 types of cancer	
Javed Khan et.al. [41]	PCA	neural networks	No combining done	cDNA microarray for tumor	Find the minimum gene set that can correctly classify the samples

Another problem that has drawn the interest of researchers working on gene classification is the class imbalance in microarray data problems. Hualong Yu et.al. [42] address the problem, and transform the multiclass to multiple binary classes, which is an evolving version of the random subspace method. Next, they attempt to correcting the class imbalance by random under sampling or decision threshold adjustment. They use SVM as a base classifier and a modified voting fusion strategy. They experiment with 8 cancer microarray datasets and propose a class insensitive classification method. Our data doesn't suffer from class imbalance before the cross validation is applied. Authors in [42] also confirm that DNA microarray data are known to contain noisy and redundant genes that must be preliminarily eliminated. They delete redundant features using Pearson correlation coefficient as a similarity measure and signal to noise measure to remove noisy genes. We use a simpler distance method of direct sample subtraction to compare features after clustering.

3. The Genome data

3.1. Data Description

Contrary to most studies that use the gene expression level where the genes are the selected features that represent the sample to be identified, in this study, we use the DNA copy number where the selected features are the genomic position which could include a gene or part of a gene.

Even though the DNA copy numbers variations occur frequently in the genome of normal people, especially in the segmental duplication regions (SDs), it has been demonstrated that some variations are associated with behavioral and developmental abnormalities such as cognitive impairment, autism, mental retardation, and possibly psychiatric diseases. Different studies tested the whole genome and detected autism-related abnormalities in five SD-rich intervals [15]. Therefore, autism is correlated with DNA copy number variations (DCV).

A case-control study has been conducted by [15, 43] using high-resolution of 1 probe/160 bp array comparative genomic hybridization (aCGH) [7 - 9] with probes covering both low copy repeats (LCRs) and surrounding sequences to evaluate 71 children with autism (AU) and 71 typically developing (TD) controls matched for ethnicity and gender. To determine if smaller, more common copy number variations (CNVs) within unstable segments of the genome contribute to autism susceptibility, five LCR-rich regions have been examined by [15, 43] where recurrent rearrangements are associated with neurobehavioral disorders. They designed a custom 385K oligonucleotide array from Roche NimbleGen Systems, Inc. (Madison, WI, USA) targeting five genomic intervals with an average probe density of one probe every 120 bp in segmental duplication-containing intervals and one probe every 200 bp in unique sequence regions.

Our study is confined to analyze and detect the recurrent variations across the five LCR-rich intervals used by [11] which have a total length of 75Mb using finely-tiled oligonucleotide arrays. The five genomic regions were chr7: 61 058 424 – 82 000 033 (20.9 Mb), chr10: 77 000 071 – 91 999 959 (15.0 Mb), chr15:18 260 026 – 34 999 973 (16.7 Mb), chr17:12 000 112 – 22 187 066 (10.2 Mb) and chr22: 14 430 001 – 26 000 041 (11.6 Mb). Segmental duplication (SD) containing regions accounted for 24.5% of the sequence on the array (18.2 out of 74.4 Mb). The 5 low-copy repeats regions with (SD)-rich are summarized in Table B1. Table B2 presents the genotype and the phenotype of neural abnormal behavior risk that include but not limited to the 5 studied regions in Table B1.

The experimental method, quantitative polymerase chain reaction (PCR) [44], has been used by [15, 43] to evaluate the association of some of the detected regions and the targeted disease. The sensitivity and specificity of the five-LCRs intervals were evaluated by comparing CNV data from two control samples with CNV data from orthogonal whole-genome platforms as reported previously by [15]. They utilize two complementary CNV detection algorithms and PCR validations and estimate a true positive

rate between 71.25 and 80% and false positive rate between 5.1 and 6.6% high-confidence set of CNVs detected with a different CNV. Details for the five-LCRs experimental methods, including platform comparisons, CNV calling criteria, subjects recruited and ascertained are provided in the Supplementary Material, Methods by [15]. The availability of the dataset used in this study is upon request by [15].

Table 2 Studied intervals of each chromosome data

chromosome	Start	end	length
7	61058424	81999980	20,941,556
10	77000071	91999901	14,999,830
15	18260026	34999924	16,739,898
17	12000112	22187009	10,186,897
22	14430001	25999992	11,569,991

3.2.Data Preprocessing

Before any feature selection and classification is performed, at the first stage, we need to clean-up the data to improve its quality using the preprocessing method. A brief description of the applied preprocessing method is presented as follows.

For a given aCGH profiles, the data can be modeled as piecewise constant autoregressive (AR) processes excited by additive white Gaussian noise (AWGN). Formally,

$$y[n]= f[n] + w[n]. \qquad n=0, 1, 2, \dots, N-1 \tag{1}$$

where $y[n]$ is the observed DCN data, $w[n]$is AWGN and $f[n]$ is the true signal to be estimated with M segments defined as

$$f[n] = \sum_{i=1}^{M} A_i \left[u[n_{i-1}] - u[n_i]\right] \tag{2}$$

where $n_0= 0< n_1< n_2< \dots < n_{M-1}< n_M= N$ and $u[n]$ is the unit step function. Here A_i and n_i are the unknown parameters representing the intensity level and the breakpoint, respectively. N is the length of data. Moreover, each variant region is assumed to be statistically independent of all other regions. Hence, the PDF of the entire data record can be written as

$$\boldsymbol{p}(\boldsymbol{y}; \boldsymbol{A}, \boldsymbol{n}) = \prod_{i=1}^{M} p_i(y[n_{i-1} : n_i - 1]; A_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{M} \left[\sum_{n=n_{i-1}}^{n_i-1} (y[n] - A_i)^2 \right] \right]. \tag{3}$$

Before further analysis, we apply a Bayesian-based estimator approach [45], to identify and detect the variant regions by discritizing the normalized aCGH datasets. The method can be summarized as follows.

1. Estimate the number of variant segments using minimum description length (MDL) [10] algorithm.

$$MDL(k) = -\ln \prod_{i=1}^{k} p_i\left(y[\hat{n}_{i-1}], \dots, y[\hat{n}_i - 1]; \hat{A}_i\right) + \frac{m_k}{2} \ln N. \tag{4}$$

where m_k is the number of estimated parameters or equivalently the dimensionality of the unknown parameters A_i's and n_i's with k breakpoints.

2. Estimate the values of the breakpoints n_i's of the variant segments maximizing the likelihood ratio test (LRT) or minimizing the least square errors.

$$\boldsymbol{J}\left(\boldsymbol{A},\ \boldsymbol{n}\right) = \sum_{i=1}^{M} \left[\sum_{n=n_{i-1}}^{n_i-1} \left(y[n] - \hat{A}_i\right)^2 \right]. \tag{5}$$

3. Evaluate the predicted segments values using the sample mean for the points within the segment boundaries.

$$\hat{A}_i = \frac{1}{n_i - n_{i-1}} \sum_{n=n_{i-1}}^{n_i - 1} y[n]. \qquad for\, i = 1:M \tag{6}$$

4. Experimental methodology:

In this paper, we present a robust method for efficient autism detection using CGH data and machine learning tools. The presented method is tackling the difficult tasks due to the large dimensionality of the data set, the high overlap in the class distributions and to the small sample size common for genetic studies. The complete autism detection process is illustrated in Figure 1. The process involves several stages; starting from the data preprocessing stage, the feature selection or reduction stage, the classifier design stage and ending at the classifier combination or decision fusion stage. After preprocessing the data using the method presented by [45], we experiment with different methods that reduce the dimensionality such as principal component analysis, PCA and clustering techniques. We experiment with nine different feature reduction methods that are derived from the two existing methods of PCA and clustering. PCA method is normally used in image processing or multispectral imaging research. Clustering method finds features that form more compact clusters with high between class separability. Different variations of clustering are examined to find the best one. At the last two stages of classification and combination, our experiments involve five classifiers and three combiner methods in addition to the single classifier. Classifiers that we experiment with are k-nearest neighbor, 1-nearest neighbor, MLP backpropagation neural networks and two types of support vector machine classifiers. Combiner methods we experiment with are bagging [46], random subspace method, 'RSM' [47], feature selection based combiner, 'FSC' [48, 49], in addition to the single classifier. In all the combiners the classifier decisions are fused using the sum soft fusion strategy [50, 51]. Our experiments involve different feature set sizes where we found the minimum size that yields an optimum performance.

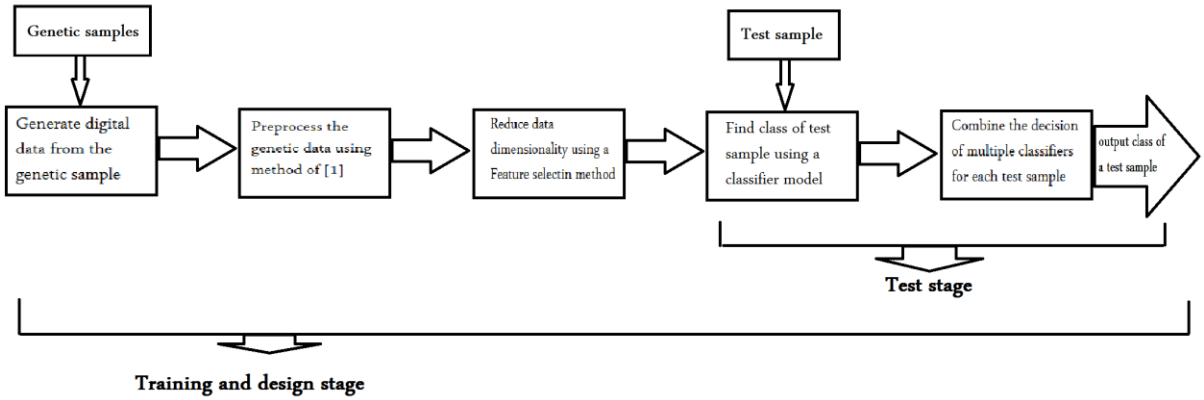

Figure 1 is a block diagram of the complete system from sampling to detection.

The success of our system on autism gene data may prompt others to use the same for other diseases and for the detection of cancer using their microarray gene expression data, in addition to setting the stage for a quicker autism detection method.

Simulation experiments are conducted using Matlab. The data is partitioned in two training and test sets based on the 10 fold cross validation method. The training set is used at the feature selection stage to select the best features. It is also used to design the classifiers and combiners while the test set is used to measure the classification rate of the system. The classification rate is found by dividing the total number of correctly classified test samples by the total number of test samples. The features used in the test set are the ones initially selected using the training set. Furthermore, the training set is divided

in two equal sets; training and validation sets. Classifier and combiner methods are validated using the validation sets.

We repeat the experiments using five classifiers as described below, and three combiners for each type of the classifiers. The fusion method used to combine the classifiers is Sum [50, 51] fusion method. Therefore, our system for autism detection consists of the following stages after conversion of genetic information found in a human sample to digital genetic data using microarray sequencing of gene expression levels.

1- Preprocessing of genetic data based on the method of [45].
2- Feature selection to reduce data dimensionality.
3- Classifier design or training.
4- Classifier combination and decision fusion.

For some combiners the third and fourth stages are merged in one step.

For the first stage we use the method of [45] as described in the previous section. The methods of stages 2 to 4 are described next.

In the results section we present boxplots that compare the rates achieved from the different methods. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually as a plus. Points are drawn as outliers if they are larger than $q3 + 1.5(q3 – q1)$ or smaller than $q1 – 1.5(q3 – q1)$, where q1 and q3 are the 25th and 75th percentiles, respectively. The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

4.1. Feature selection methods:

4.1.1. Multistage PCA:

Feature selection using the training set of each chromosome is made using PCA on Matlab through the function “princomp”. To avoid the limited memory error due to the large number of features, for each chromosome data, we apply the PCA function to blocks of 6000 features where each block returns the best d features. d is calculated as block size divided by the number of blocks. Therefore, the total features out of the first stage is equal to the block size. These features are grouped in a matrix where a second stage PCA is applied to the group of features from the first stage. These best features are found using best eigenvalues that are larger than 1, 5 and 10 to form three feature set sizes. These yield a large reduction of dimensionality that is less than 100 features. These yield gradually smaller number of features referred to in the table of results as eigen1, eigen2 and eigen3, respectively. Additionally a fourth set which is the largest set and includes 400 features is created based on the largest eigen values.

When selecting features in a block from the full set we experiment with two procedures, serial and random. Using the serial method we take the 6000 features in sequence from the fixed list of features until the required number of features is reached. Using the random method the required number of features are taken randomly, without replacement, from the full set of features.

Table 3 number of features experimented for each feature selection type and set size, V stands for variable number of features.

Feat. Set size	sPCA	rPCA	Clustering	PCA - cluster	Cluster-PCA	2nd stage clustering	3rd stage clustering	4th stage clustering	Cluster-Del
1	V	V	50	50	V	50	50	50	50
2	V	V	30	30	V	30	30	30	30
3	V	V	10	10	V	10	10	10	10
4	400	400	400	100	400	-	-	-	V < 400

Table 4 Number of features at the three feature set sizes found using serial PCA

Chromosome	Eigen 1	Eigen 2	Eigen 3
chr7	68	12	5
chr10	27	6	2
chr15	83	18	11
chr17	44	6	4
chr22	66	16	8

4.1.2.Clustering

We don't aim to find classes or clusters because the classes are known. However, we aim to use clustering tools to find the most distinguishing features. Therefore, we attempt to use clustering tools to find features that yield the largest distance between the means of the two clusters and yield clusters with smallest standard deviation. This can be found using the following equation, known as fisher score [52].

$$f_{1,2} = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 - \sigma_2)} \quad (7)$$

Based on their fisher scores we sort the features in a descending order. We experiment with taking the best 50, 30 and 10 features, which are referred to as size 1, 2 and 3 in the tables of results. Additionally we experiment with a fourth very large size of the best 400 features.

4.1.3.PCA-Clustering

This method is similar to the two stage PCA, however at the second stage we use clustering to find the best features instead of PCA. Again we select the best 50, 30 and 10 features from the sorted list. The largest fourth size is set to include 100 features, due to the small number of features passing the first stage.

4.1.4.Clustering-PCA:

Here we sort features according to the clustering method which used the fisher score equation 7. Then apply PCA to the best features to obtain a new representation of the feature space where features are moved to a more distinguishing representation. Best eigenvectors are found by finding eigenvalues that are greater than 0.01, 0.1 and 0.5. The fourth set is the largest set that includes 400 features from the sorted list of features with highest eigen values.

4.1.5.Staged Clustering, "2nd, 3rd, 4th":

This is a feature selection method that is proposed by us and is based on clustering but we take the most different features by measuring the Euclidean distance between features. The furthest 1000 are taken and clustering is applied to them. Next, from this sorted list the furthest 500 are taken and clustering is applied to them to create the 2nd stage sorted cluster set. Next for this sorted list the furthest 100 are taken and clustering is applied to them to create the 3rd stage cluster set. The 4th stage cluster set is created by taking the furthest 50. These are referred to as 2nd stage, 3rd stage and 4th stage. For each of these three feature selection methods we consider three feature set sizes that are used by the classification system. The feature set sizes are 50, 30 and 10 features, referred to as size 1, 2 and 3 respectively. No fourth large set size exists for this method.

4.1.6.Clustering with deleted redundancies, "Cluster-Del":

This Feature selection method finds the best features according to the clustering method of equation 7, then sorts them. The top 10000 features with the highest fisher score are considered for further processing where the similar features are deleted. The remaining number of features varies. For example for chromosome chr7 at cross validation 1 it is 1032. Indicating a high redundancy of features in the

data. The process of fisher clustering is repeated again on the remaining features and the top 500 features are considered for deletion of similar features. For example for chromosome chr7 at cross validation 1 the remaining number of features after removal of similar features is 297. Note that we may have similar features at the second stage which were not removed at the first stage. This is due to the redundancy removal algorithm which skips a feature if two consecutive ones are similar to the original feature under investigation. Whether redundancies are removed at the first stage or at two stages does not affect the final outcome of features. Similar to the rest of the feature selection methods we also experiment with three other feature set sizes created by taking the top 50, 30, and 10 features from the sorted and deleted list of the second stage.

Table 3 presents the number of features used for each feature selection method at the four sizes under investigation. V stands for a variable number of features for each cross validation and chromosome type. For sPCA for example at cross validation 1 we obtain the number of features displayed in Table 4

4.2. Classifier types

For the nearest neighbor classifier we experiment with two values of k set at 1 and $\sqrt{N}$, where N is the square root of the number of training samples. The distance metric used is the mahalanobis metric. The neural network classifier used here consists of three layers. The transfer function or output of the first two layers is log-sigmoid, while that of the output or third layer is purelin, see Figure *2*. The network training function used is backpropagation. The number of neurons in the first layer is equal to the number of features, while that for the hidden (second) layer is set at 5. The number of neurons at the output layer is equal to the number of classes, which is two. For the support vector machine, SVM [53], we experiment with two SVMs; one with RBF sigma and box constraint values set to 1, and a second with these parameter values calculated using the training set and set to 0.3.

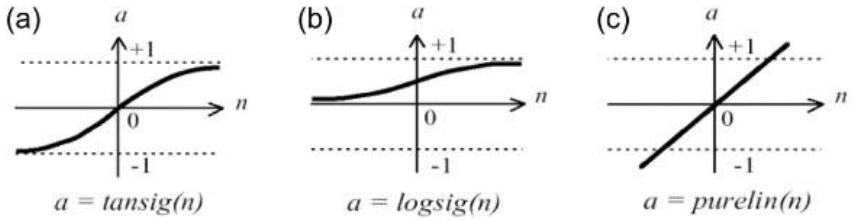

Figure 2 types of neural activation functions

4.3. Combiner systems:

Bagging predictors proposed by Breiman [46], is a method of generating multiple versions of a predictor or classifier, via bootstraping and then using those to get an aggregated classifier. We set the number of multiple versions of classifiers to 25, as recommended by Breiman [46]. The total number of samples in each bootstrap set is equal to those of the original training set. The second combiner 'RSM' [47] aims at creating diverse classifiers by assigning different features to each classifier. The number of features is set at a fixed value, m, less than the total number of features. Each classifier is assigned a subset of features that are randomly selected without replacement from the full feature set. This results in classifiers having different views of the data space. We set m to equal 67 percent of the total number of available features. In comparison to 50% recommended by [47] we found better rates are achieved at 67%. The number of combined classifiers is set equal to bagging at 25.

The third combiner is the feature selection based combiner, FSC, proposed by Alkoot & Kittler [48, 49] and it is based on the principal that the feature selection and the combiner performance are linked. The best feature subset is selected for each classifier based on the combiner system performance instead of the individual classifier performance. The maximum possible number of classifiers that can be fused in

the system is limited by the number of available features. We have set it to 5 maximum number of classifiers. Any feature selection method can be used to add the best feature subset such that the combiner system error rate is minimized. For each classifier under construction one feature is inserted at a time and the system performance is checked. After checking all features the feature yielding the best system performance is inserted to the classifier under construction. When the feature insertion process is completed for the maximum number of classifiers in the system the process is repeated from the first classifier until all features are used up or the system error rate is not improved by the insertion. The process continues as long as the addition of a new feature does not degrade the system performance, and there are an unused number of features. However, on the first run across the classifiers we add a feature even if it does not improve the system. That is we force the insertion of the best feature to the classifiers, even if that does not improve the system. The feature selection method used is the 2-forward-1-backward method.

5. Results:

We are interested in monitoring the performance of the different methods at each stage separately. Comparing the results of the feature selection methods, at the second stage, given the different classifiers and combiners indicates that the three PCA based methods yield lower rates that are mostly in the 60's. We also noticed that increasing the number of features, by reducing the eigen value of acceptable features, does not yield a significant improvement of results. For the clustering based feature selection methods we found that rates are mostly in the 70's and occasionally in the 80's. The increase in the number of feature set size yields an improvement in the performance.

Table 5 Best classification rate achieved by each Feature selection method

Chromo	Eigen	Cluster-Del	Clstr	2nd stage	3rd stage	4th stage	Clstr-PCA	PCA-Clstr	sPCA	rPCA
chr7	1	**77.86**	70.71	70	64.29	65.71	74.29	63.57	68.57	69.29
	2	**73.57**	69.29	70	60.71	65.71	68.57	66.43	70	72.14
	3	69.29	70	70	58.57	67.14	67.14	68.57	**76.43**	**75.71**
	4	**100**	72.86	-	-	-	69.29	65	72.14	70
chr10	1	65	**70.71**	57.86	67.86	67.86	70.71	64.29	67.14	67.86
	2	65	71.43	58.57	66.43	68.57	**72.14**	57.86	64.29	65
	3	61.43	**70.71**	60	68.57	68.57	64.29	56.43	56.43	62.86
	4	**100**	70	-	-	-	66.43	59.29	67.86	68.57
chr15	1	80.71	75.71	77.14	**82.86**	**82.86**	**82.86**	70	68.57	70
	2	78.57	75	75	**80**	**80**	79.29	63.57	64.29	64.29
	3	**83.57**	74.29	66.43	77.14	79.29	67.86	60	63.57	62.14
	4	**100**	81.43	-	-	-	73.57	67.14	60.71	56.43
chr17	1	79.29	82.14	**87.14**	76.43	74.29	73.57	69.29	65.71	67.86
	2	78.57	80	**85.**	74.29	74.29	73.57	65	67.14	65.71
	3	75.71	80	**83.57**	73.57	73.57	65.71	58.57	67.14	67.14
	4	**99.29**	86.43	-	-	-	74.29	65.71	70	68.57
chr22	1	68.57	**76.43**	71.43	68.57	71.43	71.43	64.29	67.86	69.29
	2	65.71	**74.29**	71.43	68.57	70.71	70.71	65	68.57	67.86
	3	67.14	71.43	**76.43**	68.57	67.14	67.86	60.71	70	72.14
	4	**99.29**	75.71	-	-	-	69.29	63.57	72.14	68.57
chr all	1	**80**	77.86	76.43	77.14	77.14	76.43	71.43	65.71	66.43
	2	**79.29**	77.86	74.29	76.43	76.43	72.86	66.43	69.29	64.29
	3	74.29	**77.14**	70.71	71.43	71.43	69.29	67.86	65	62.86
	4	**98.57**	90	-	-	-	76.43	70	69.29	66.43

We found that the "Cluster-Del" method outperforms all other methods at the largest feature set size reaching the optimum 100% or closely lower, for all chromosomes. At lower feature set sizes lower

rates are achieved where different feature selection methods rank top at the different chromosomes. "2nd stage clustering" method is best at chr17 and smallest feature set size of chr22. Regular clustering is best at chr10 and chr22. 3rd stage, 4th stage and cluster-PCA are best at chr15 sizes 1 and 2. PCA is best only at the smallest feature set size of chr7. At the rest of the sizes and chromosomes Cluster-Del is best.***Error! Reference source not found.*** shows the maximum rate achieved by each of the feature s election methods at the different chromosomes and feature set sizes. **Error! Reference source not found.** shows the performance of the Cluster-Del feature selection method and its difference with the other methods at all sizes. Table *7* presents all rates that are insignificantly lower than the maximum. Calculation of significance is made by finding rates that are lower by less than five percent of the amount needed for the maximum rate to reach perfect classification, i.e. 100%. This can be found through an equation which finds the lowest classification rate considered insignificantly lower than the highest rate achieved,

Closest lower acceptable classification rate = Highest rate – (5% x (100-highest rate))

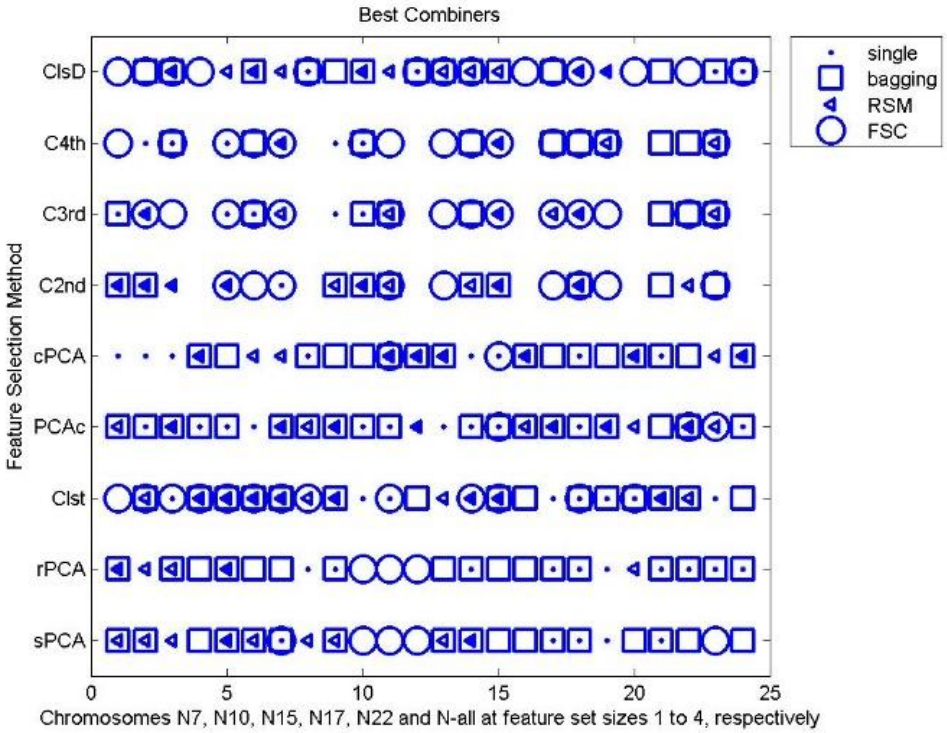

Figure 3 display of best combiner at each chromosome size and feature selection method.

At the third stage, we compare classifier performances, by looking at Figure *8* and Table *7*. We found that the best achieved rate using Cluster-Del feature selection method at the largest feature set size, was using the neural network classifier combined using FSC, except at chromosome chr15. kNN and 1-NN occasionally achieved the maximum rate at chr10 and chr15, when combined using bagging or FSC. All three classifiers, neural network, kNN and 1NN, yielded the maximum rate using FSC for the N-all data. For this data 1-NN also achieved the best rate using bagging.

Table 6 Difference in classification rate between the cluster with deletion, Cluster-Del, feature selection method and the eight other feature selection methods. Negative values indicates how much the Cluster-Del yielded a lower classification rate.

Chromo	Eigen	**Cluster-Del**	Clstr	2nd stage	3rd stage	4th stage	Clstr-PCA	PCA-Clstr	sPCA	rPCA
chr7	1	**77.86**	7.15	7.86	13.57	12.15	3.57	14.29	9.29	8.57
	2	**76.43**	4.28	3.57	12.86	7.86	5.	7.14	3.57	1.43
	3	72.14	**-0.71**	**-0.71**	10.72	2.15	2.15	0.72	**-7.14**	**-6.42**
	4	**100.00**	27.14				30.71	35.	27.86	30.
chr10	1	65.71	**-5.71**	7.14	**-2.86**	**-2.86**	**-5.71**	0.71	**-2.14**	**-2.86**
	2	66.43	**-6.43**	6.43	**-1.43**	**-3.57**	**-7.14**	7.14	0.71	0
	3	62.14	**-9.28**	1.43	**-7.14**	**-7.14**	**-2.86**	5.	5.	**-1.43**
	4	**100.00**	30.				33.57	40.71	32.14	31.43
chr15	1	81.43	5.	3.57	**-2.15**	**-2.15**	**-2.15**	10.71	12.14	10.71
	2	79.29	3.57	3.57	**-1.43**	**-1.43**	**-0.72**	15.	14.28	14.28
	3	**83.57**	9.28	17.14	6.43	4.28	15.71	23.57	20.	21.43
	4	**100.00**	18.57				26.43	32.86	39.29	43.57
chr17	1	80.71	**-2.85**	**-7.85**	2.86	5.	5.72	10.	13.58	11.43
	2	81.43	**-1.43**	**-6.43**	4.28	4.28	5.	13.57	11.43	12.86
	3	75.71	**-4.29**	**-7.86**	2.14	2.14	10.	17.14	8.57	8.57
	4	**99.29**	12.86				25.	33.58	29.29	30.72
chr22	1	70.00	**-7.86**	**-2.86**	0	**-2.86**	**-2.86**	4.28	0.71	**-0.72**
	2	65.71	**-8.58**	**-5.72**	**-2.86**	**-5.**	**-5.**	0.71	**-2.86**	**-2.15**
	3	67.14	**-4.29**	**-9.29**	**-1.43**	0	**-0.72**	6.43	**-2.86**	**-5.**
	4	**99.29**	23.58				30.	35.72	27.15	30.72
chr all	1	**81.43**	2.14	3.57	2.86	2.86	3.57	8.57	14.29	13.57
	2	**79.29**	1.43	5.	2.86	2.86	6.43	12.86	10.	15.
	3	74.29	**-2.85**	3.58	2.86	2.86	5.	6.43	9.29	11.43
	4	**98.57**	8.57				22.14	28.57	29.28	32.14

Comparing the results of combiners at the best feature selection method of Cluster-Del, at the largest feature size, we found that the best are FSC and occasionally bagging, while RSM always underperformed. For all the data sets, at the largest feature set size, FSC using neural networks was consistently the best combiner. Except at chr15 where FSC was best using kNN and 1-NN, while neural networks was closely behind. Bagging was equal to FSC using kNN and 1-NN at two data sets only, chr10 and chr15. Single was also equal to FSC using kNN and 1-NN at chr10, chr15 and N-all. In figures 4 to 7 boxplots show the performance of each combiner at the various chromosomes and feature set sizes. We find that FSC yields the lowest variance in results compared to other combiners indicating robustness. Contrary to other combiners, it never reaches below 50 percent. Also we found that at the largest feature set size 4, minimum rates recorded, using FSC with other classifiers, are higher than the minimum rates of the other combiners. In Table 7 we also find that for the N-all data that is a merge of all chromosomes, FSC yields the maximum rate using any of the three classifier types; neural, kNN or 1NN.

Comparing results of the system consisting of the clustering-Del, FSC and neural network for the different chromosome data sets, we found that the optimum rate of 100% was achieved for most chromosomes. At the largest feature set size chr17 and chr22 yielded lower rates of 99.29, while the merging of all chromosomes, i.e. N-all, yielded a further lower rate of 98.57. These rates are higher than other sizes or feature selection methods. Looking at the statistics of each chromosome in Figure 9, we see the classification rates averaged over the different feature selection methods, classifiers and combiners, and find three groups of performances. We find that chromosome 10 yields the lowest

classification rate averaged over the different classifier and combiner types. Chromosomes 7 and 22 yield higher rates and variances. Chromosomes 15 and 17 yield the highest average rates with a higher variance. The merging of all chromosomes yielded average rates and variances that are an average between the three groups. Figure 10 displays a comparative image that shows the maximum rate achieved by each feature selection method at the different chromosome sizes. It indicates that PCA methods are generally yielding lower rates. It also shows that chr10 chromosome yields the most number of low rates while chr15 and chr17 yield the largest number of high rates. The figure also displays the largest size at which an optimum rate is achieved by each of the chromosomes.

Table 7 clustering with deleted redundancies, "Cluster-Del".

Chromosome	Eigen	Combiner system	classifier	Classification rate
chr7	1	FSC	SVM.3	77.86
	2	Bagging FSC FSC	N. Net SVM1 SVM.3	73.57 72.86 72.14
	3	Single FSC, bagging, RSM bagging	N.Net N.Net k-NN	69.29 67.86 67.86
	4	FSC	N.Net	**100**
chr10	1	RSM	N.Net	65
	2	RSM Single, bagging	N.Net	65 63.57
	3	RSM	N.Net, SVM1	61.43
	4	FSC Single, bagging, FSC	N.Net kNN, 1-NN	**100** **100**
chr15	1	Bagging	kNN	80.71
	2	Single, bagging, bagging	kNN N.Net	78.57 78.57
	3	RSM	SVM1	83.57
	4	Single, bagging, FSC	kNN, 1-NN	**100**
chr17	1	FSC, RSM FSC bagging	N.Net kNN N.Net	79.29 78.57 78.57
	2	FSC, RSM bagging	N.Net	78.57 77.86
	3	Bagging RSM	N.Net	75.71 75
	4	FSC	N.Net	**99.29**
chr22	1	FSC, Bagging FSC	N.Net kNN	68.57 68.57
	2	Single FSC, RSM	kNN kNN	65.71 65
	3	RSM Single	SVM1	67.14 65.71
	4	FSC	N.Net	**99.29**
chr all	1	Bagging	N.Net	80
	2	FSC	kNN	79.29
	3	Single	N.Net	74.29
	4	Single, FSC Single, bagging, FSC FSC	kNN 1-NN N.Net	**98.57**

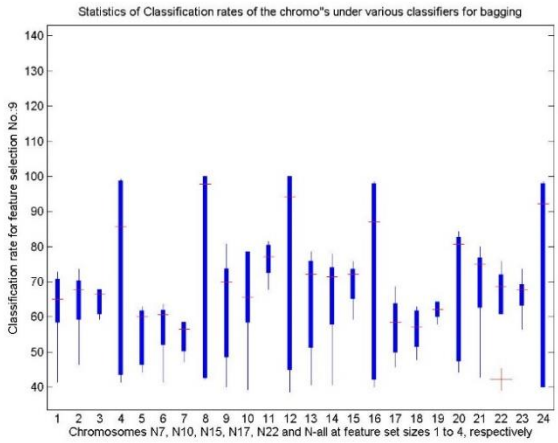

Figure 4 Statistics of the bagging combiner classification rate at different chromosomes using different classifiers

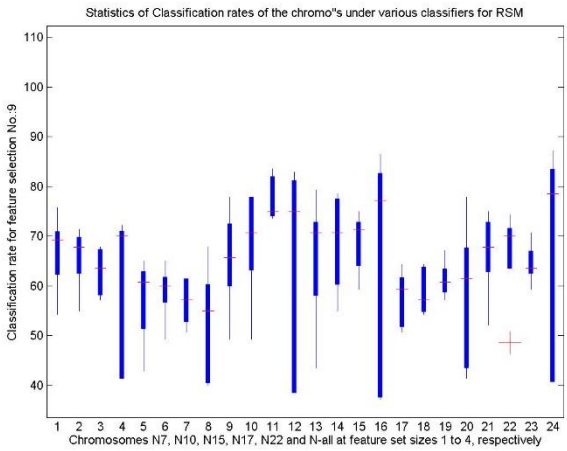

Figure 6 Statistics of the RSM combiner classification rate at different chromosomes using different classifiers

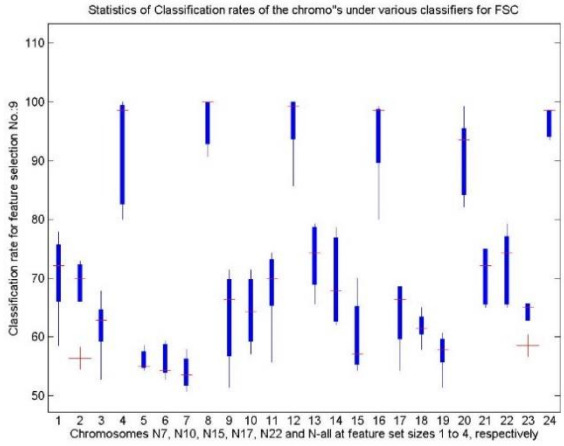

Figure 5 Statistics of the FSC combiner classification rate at different chromosomes using different classifiers

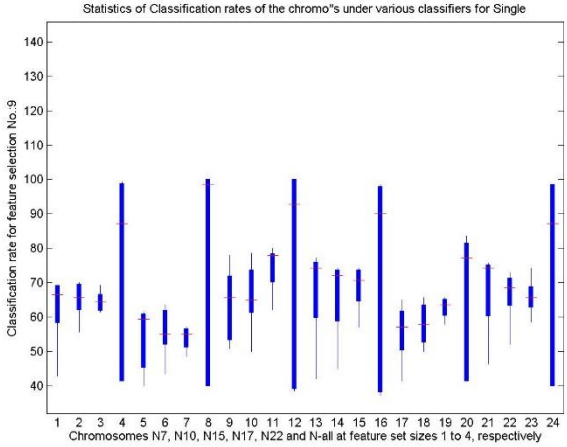

Figure 7 Statistics of the Single classifiers classification rate at different chromosomes using different classifiers

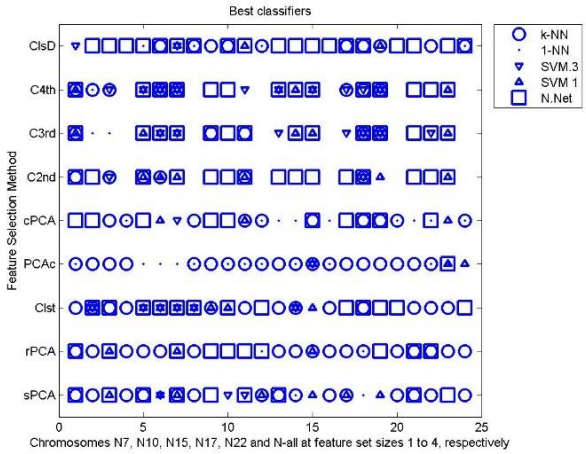

Figure 8 display of best classifiers at the different chromosome sizes and feature selection methods. Clustering methods C2nd, C3rd and C4th were not experimented at chromosome size 4

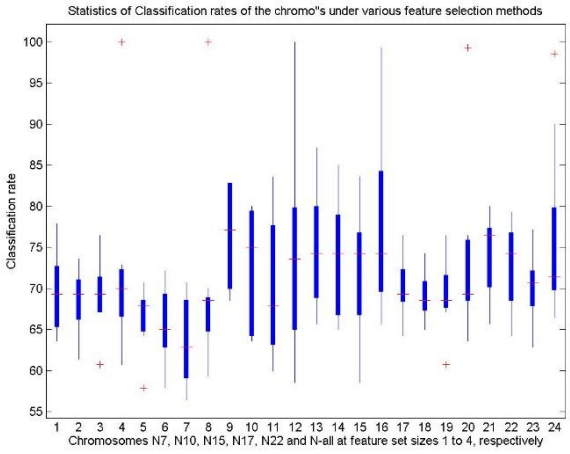

Figure 9 Statistics at different chromosomes over different combiners, classifiers, and feature selection methods

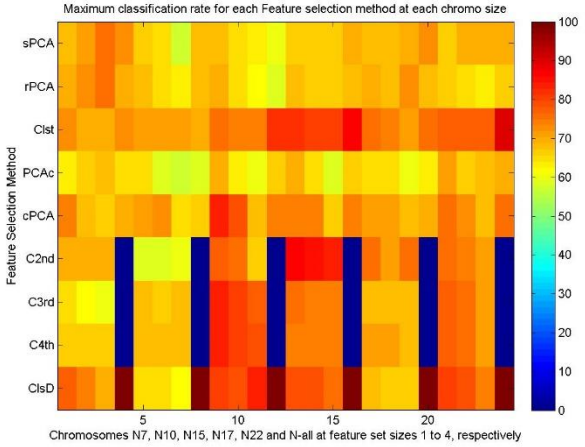

Figure 10 maximum classification rate of nine feature selection methods at each chromosome size. (sPCA: serial PCA, rPCA: random PCA, clst: Clustering, PCAc: PCA-Clustering, cPCA: clustering-PCA, C2nd: 2nd stage sorted clustering, C3rd: 3rd stage sorted clustering, C4th: 4th stage sorted clustering, ClsD: Clustering with deleted redundancies)

6. Conclusion

Autism is a disorder with detrimental effects that increase with age. This effect can be reduced if the disorder is detected early and therapy is introduced. Current detection methods are based on behavioral examinations which lead to delayed detection. Using machine learning techniques we aim to use CGH data obtained from five chromosomes to design an automated system that speeds up the detection process. The system helps to improve detection, identification and diagnosis of autism. This will benefit both victims and society in general and will lead to early diagnosis and new treatments.

The designed system consists of four stages of preprocessing, feature selection (or dimensionality reduction), classification then classifier combination. The importance of the second stage is due to the widely known curse of dimensionality. For the data under investigation, without dimensionality reduction, classifier decisions would be similar to a wild guess. We experiment with existing methods such as PCA and clustering to reduce the data dimensionality. We proposed several variants of these methods that lead to significant improvements over existing methods. For the third stage we experimented with k-nearest neighbor, 1-nearest neighbor, backpropagation neural network and support vector machine classifiers. At the fourth stage combiner methods used were bagging, random subspace, (RSM) and feature selection based combiners, (FSC). Sum fusion was used to fuse the component classifier decisions.

We repeated the experiments at four feature set sizes and found that optimum performance can be achieved using the largest feature set size, which includes a number of features between 150 and 500 features, depending on the chromosome data set. This optimum 100% or closely lower rate was achieved using the neural network classifier when combined using FSC and only when clustering with deleted redundancies feature selection method was used. While FSC yields the best rate for all chromosomes, bagging with nearest neighbor classifiers yielded this optimum rate for two of the chromosomes.

The application of machine learning tools for identification of autism using CGH data is rare. Through our proposed technique we showed that it is possible to detect autism using CGH data through machine learning techniques. The implementation of such a system will lead to early intervention and enables us to detect if a subject has the potential to develop autism using the subjects' gene data, even before any behavioral symptoms start to appear.

References

1. Lauritsen, M. and Ewald, H. (2001), The genetics of autism. Acta Psychiatrica Scandinavica, 103: 411–427. doi: 10.1034/j.1600-0447.2001.00086.x
2. Yonan, A. L., Palmer, A. A., Smith, K. C., Feldman, I., Lee, H. K., Yonan, J. M., Fischer, S. G., Pavlidis, P. and Gilliam, T. C. (2003), Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. Genes, Brain and Behavior, 2: 303–320. doi: 10.1034/j.1601-183X.2003.00041.x
3. A. Bailey, A. LeCouteur, I Gottesman, P. Bolton, E. Simonoff, E. Yuzda; M. Rutter .Autism as a strongly genetic disorder: evidence from a British twin study. Psychol Med. 25:63-77. 1995.
4. Lord C1, Risi S, Lambrecht L, Cook EH Jr, Leventhal BL, DiLavore PC, Pickles A, Rutter M. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum ofautism. J Autism Dev Disord. 2000 Jun;30(3):205-23.
5. Dr. Paul T. Shattuck, Maureen Durkin, Matthew Maenner, Craig Newschaffer, David S. Mandell, Lisa Wiggins, Li-Ching Lee, Catherine Rice, Ellen Giarelli, Russell Kirby, Jon Baio, Jennifer Pinto-Martin, and Christopher Cuniff. The Timing of Identification among Children with an Autism Spectrum Disorder: Findings from a Population-Based Surveillance Study. J Am Acad Child Adolesc Psychiatry. 2009 May; 48(5): 474–483.
6. Wiggins LD1, Baio J, Rice C. Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample. J Dev Behav Pediatr. 2006 Apr;27(2 Suppl):S79-87.

7. Kallioniemi A, Kallioniemi OP, Sudar Da, Rutovitz D, Gray JW, Waldman F, Pinkel D "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," Science, 258:818-821, 1992.
8. Weiss M, Hermsen M, Meijer G, Van Grieken N, Baak J, Kuipers E, Van Diest P (1999) Comparative genomic hybridization. Molecular Pathology 52:243-251.
9. Pinkel D, Albertson DG "Comparative genomic hybridization." Annu Rev Genomics Hum Genet, Vol. 6, p 331-354, 2005.
10. Torsten Rohlfing, Daniel Russakoff and Calvin Maurer, Performance based classifier combination in atlas based image segmentation using expectation maximization parameter estimation, IEEE transaction on medical imaging, vol 23, no 8, August 2004.
11. Inan Guler and Elif Derya Ubeyli, ECG beat classifier designed by combined neural network model, Pattern Recognition, 38 (2005) 199-208.
12. M.P.Sampat, et. al., Supervised parametric and non parametric classification of chromosome images, Pattern Recognition 38(2005) 1209-1223.
13. Alexey Tsymbal , Padraig Cunningham, Mycola Pechenizkiy and seppo Puuronen, Search strategies for ensemble feature selection in medical diagnosis, 16th IEEE symposium on computer based medical systems, 2003, June 2003, 124-129.
14. Hyunseok kook et. al., Multi-stimuli multi-channel data and decision fusion strategies for dyslexia prediction using neonatal ERPS, Pattern Recognition vol 38, no 11, 2005, 2174-2184.
15. Girirajan, S., Johnson, R. L., Tassone, F., Balciuniene, J., Katiyar, N., Fox, K., Selleck, S. B.. "Global increases in both common and rare copy number load associated with autism. Human Molecular Genetics," Vol. 22, No. 14, pp. 2870–2880, 2013.
16. DP Wall, J. Kosmicki, TF DeLuca, E Harstad and VA Fusaro. Use of machine learning to shorten observation based screening and diagnosis of autism. Transi Psychiatry, 2,e100. 2012.
17. Mohammed Uddin, Kristiina Tammimies, Giovanna Pellecchia, Babak Alipanahi, Pingzhao Hu, Zhuozhi Wang, Dalila Pinto, Lynette Lau, Thomas Nalpathamkalam, Christian R Marshall, Benjamin J Blencowe, Brendan J Frey, Daniele Merico, Ryan K C Yuen, & Stephen W Scherer, Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. Nature Genetics, Vol. 46, Pp: 742–747 :(2014).
18. A.-L. Boulesteix, C. Strobl, T. Augustin and M. Daumer. Evaluating microarray based classifiers: an overview. Cancer Informatics, 6:77-97, 2008.
19. Tomasz Latkowski, and Stanislaw Osowski. "Computerized system for recognition of autism on the basis of gene expression microarray data," Computers in Biology and Medicine, Vol: 56, Issue C, January 2015, Pages 82-88.
20. Tomasz Latkowski, and Stanislaw Osowski. "Data mining for feature selection in gene expression autism data, Expert systems with Applications," Volume 42, Issue 2, 1 February 2015, Pages 864–872
21. Elena Kostadinova, Veselka Boeva and Niklas Lavesson. Clustering of Multiple microarray experiments using information integration. C.Bohm et al (Eds.): ITBAM 2011, LNCS 6865, pp. 123-137.
22. Veselka Boeva, Anna Hristoskova and Elena Tsiporkova. Clustering of multiple microarrays through combination of particle swarm intelligence and k-means.
23. Jung Choi, Ungsik Yu, Sangsoo Kim and Ook Joon. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics. Vol.19, suppl 1, pp 184-190. 2003.
24. Anna Hristoskova, Veselka Boeva and Elena Tsiporkova. A formal Concept analysis approach to consensus clustering of multi-experiment expression data. BMC Bioinformatics, 15:1, 2014.
25. Yves Moreau , Frank De Smet , Gert Thijs , Kathleen Marchal , Bart De Moor. Functional bioinformatics of microarray data: from expression to regulation. Proceedings of the IEEE, Volume:90 Issue:11.
26. Kyung-joong Kim, Sung-Bae Cho, An evolutionary Algorithm Approach to optimal ensemble classifiers for DNA microarray data analysis. IEEE trans. On Evolutionary computation, vol 12, no 3, 2008.
27. Sung-bae Cho , Hong-hee Won. Data mining for gene expression profiles from DNA microarray. International Journal of Software Engineering and Knowledge Engineering. 13:593-608, 2003.
28. S.-B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," Proc. IEEE, vol.90, no. 11, pp. 1744–1753, 2002.

29. S. Dudoit, J. Fridlyand, and P Speed. Comparison of discrimination methods for classification of tumers using gene expression data. J. American statistics Association. Vol.97, pp77-87, 2002.
30. J.W.Lee J.B Lee, M. Park, and S.H. Song. An extensive comparison of recent classification tools applied to microarray data. Computational statistics and Data analysis. Vol 48, pp869-885. 2005.
31. A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification" Appl. Bioinformatics, vol. 2, pp. s75–s83, 2003.
32. M. Dettling, "Bagboosting for tumor classification with gene expression data," Bioinformatics, vol. 20, no. 18, pp. 3583–3593, 2004.
33. M. Dettling and P. Buhlmann, "Boosting for tumor classification with gene expression data," Bioinformatics, vol. 19, no. 9, pp. 1061–1069, 2003.
34. G. Valentini, M. Muselli, and F. Ruffino, "Cancer recognition with bagged ensembles of support vector machines," Neurocomputing, vol. 56, pp. 461–466, 2004.
35. X. Xu and A. Zhang, "Boost feature selection: A new gene selection algorithm for microarray dataset," in Proc. 6th Int. Conf. Comput. Sci.:Workshop Bioinformatics Res. Appl., 2006, pp. 670–677.
36. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, Vol. 286 no. 5439 pp. 531-537, 1999.
37. Loris Nanni, Sheryl Brahnam http://bioinformatics.oxfordjournals.org/content/28/8/1151.full - aff-1and Alessandra Lumini. Combining multiple approaches for gene microarray classification. Bioinformatics, Volume 28, Issue 8, Pp. 1151-1157
38. Yuehui Chen, Author Vitae, Yaou Zhao, Yuehui Chen, Author Vitae, Yaou Zhao, Yuehui Chen, Yaou Zhao. A novel ensemble of classifiers for microarray data classification. Applied Soft Computing Volume 8, Issue 4, Pages 1664–1669. 2008
39. Sung-Bae Cho and Jungwon Ryu Classifying gene expression data for cancer using classifier ensembles with mutually exclusive features. Proceedings of the IEEE, Volume:90 Issue:11, 2002.
40. Sung-bae Cho , Hong-hee Won . Data mining for gene expression profiles from DNA microarray. Physiol Genomics. 16;25(3):355-63. 2006.
41. Javed Khan, Jun S. Wei, Markus Ringnér, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson & Paul S. Meltzer . Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7, 673 - 679 (2001)
42. Hualong Yu, Shufang Hong, Xibei Yang, Jun Ni, Yuanyuan Dan, and Bin Qin Recognition of multiple imbalanced cancer types based on DNA microarray data using ensemble classifiers. BioMed Research International, Volume 2013 (2013)
43. Jorune Balciuniene, et al., "Recurrent 10q22-q23 Deletions: A Genomic Disorder on 10q Associated with Cognitive and Behavioral Abnormalities,"The American Journal of Human Genetics, Vol. 80, pp. 938 – 947, 2007.
44. Tassone, F., Pan, R., Amiri, K., Taylor, A.K. and Hagerman, P.J. "A rapid polymerase chain reaction-based screening method for identification of all expanded alleles of the fragile X (FMR1) gene in newborn and high-risk populations. " J. Mol. Diagn., 10, 43 – 49, 2008.
45. Abdullah Alqallaf and Ahmed Tewfik, "Maximum Likelihood Principle for DNA Copy Number Analysis," IEEE Int'l Conference on Acoustics, Speech, and Signal Processing, IEEE/ICASSP, Taipei, Taiwan, April, 2009.
46. L. Breiman. Bagging predictors. Machine Learning, 24:123–140, 1996.
47. T. Ho. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8):832{844, 1998.
48. F. M. Alkoot and J. Kittler. Feature selection for an ensemble of classifiers. In Proceedings of the SCI 2000conference, pages 622--627, Orlando, Florida, 2000 .
49. F. M. Alkoot and J. Kittler. Multiple expert system design by combined feature selection and probability level fusion. In Proceedings of the Fusion 2000 conference, volume II, pages ThC5(9--16), Paris, France, 2000 .

50. F. M. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. Pattern Recognition Letters, 20(11-13):1361–1369, 1999.
51. J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(3):226–239, 1998.
52. A.R. Webb. Statistical pattern recognition, 2nd ed., John Wiley and sons. 2002.
53. C. Cortes and V Vapnik. "Support-vector networks". Machine Learning 20 (3): 273, 1995.
54. Kimberly A. Aldinger, Jasmine T. Plummer, Shenfeng Qiu, Pat Levitt. "Genetics of Autism," Neuron, Vol. 72, Issue 2, p418–418, October 2011.
55. Stephan J. Sanders et al., "Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism," Vol. 70, Issue 5, p863–885, June 2011.

56. Hojin Moon, Hongshik Ahn, Ralph L Kodell, Chien-Ju Lin, Songjoon Baek, and James J Chen, Classification methods for the development of genomic signatures from high-dimensional data, Genome Biol. 2006; 7(12): R121.
57. D. W. Hosmer and S. Lemeshow. Applied Logistic Regression. John Wiley and sons, 1989.
58. Vlaidmir Filkov and Steven Skiena. Integrating microarray data by consensus clustering.
59. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. et al. "Global variation in copy number in the human genome." Nature, Vol. 444, pp. 444 – 454, 2006.
60. Bregje WM van Bon, et al., "The phenotype of recurrent 10q22q23 deletions and duplications," European Journal of Human Genetics, Vol. 19, pp. 400 – 408, 2011.
61. Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M. et al. "Relative burden of large CNVs on a range of neurodevelopmental phenotypes," PLoS Genet., Vol. 7, e1002334, 2011.
62. Hertz-Picciotto, I., Croen, L.A., Hansen, R., Jones, C.R., van de Water, J. and Pessah, I.N. "The CHARGE study: an epidemiologic investigation of genetic and environmental factors contributing to autism. Environ," Health Perspect., Vol. 114, 1119–1125, 2006.
63. Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L. and Spence, S.J. "The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric neuropeptide conditions," Am. J. Hum. Genet., Vol. 69, 463–466, 2001.

Appendix A

A.1 Principal Component Analysis, PCA:

PCA is a method used to extract useful information in data using statistical techniques. This yields a rearrangement of the feature space to highlight features with most information. Based on PCA the steps required to obtain a new representation of the data are:

1- Zero the mean of the data along each dimension by subtracting the mean from each dimension.
2- Calculate the covariance matrix. For an n dimension data this will be an n by n matrix. Therefore, for our data that has a large number of features, the process must be applied, in a multistage PCA, to small groups of features, in parallel, then applied again to the best outcomes of the small groups of features.
3- Calculate the eigenvectors and eigenvalues of the covariance matrix.
4- The eigenvectors of the highest eigenvalues are the principal components or the required new feature space with higher information.

A.2 Clustering

Data clustering is commonly used to find clusters, or classes, of data in an unsupervised classification problem. All methods start by defining a temporary cluster center that is gradually moved as relevant samples are assigned to the cluster. The methods differ in the techniques used to assign samples to clusters. Additionally, several methods are used to merge or divide clusters.

Appendix B

Table B2 Chromosomal regions and genes that are implicated in risk for ASD, and associated genetic disorders and syndromes including the 5 targeted regions of SD-rich described in Table B1 [54, 55].

	Chromosome region	Gene	Phenotype
Mendelian Syndromes	6q23.3	AHI1	Joubert syndrome
	7q35-q36.1	CNTNAP2	Recessive EPI syndrome, ASD, ADHD, TS, OCD
	9q34.13	TSC1	Tuberous Sclerosis type I
	10q23.31	PTEN	Cowden disease*
	11q13.4	DHCR7	Smith-Lemli-Opitz syndrome
	12p13.33	CACNA1C	Timothy syndrome
	15q11.2	UBE3A	Angelman syndrome
	16p13.3	TSC2	Tuberous Sclerosis type II
	17q11.2	NF1	Neurofibromatosis
	Xp21.2	DMD	Duchenne muscular dystrophy
	Xp21.3	ARX	LIS, XLID, EPI, ASD
	Xp22.13	CDKL5	X-linked infantile spasm syndrome
	Xq27.3	FMR1	Fragile X syndrome
	Xq28	MECP2	Rett syndrome
Rare Variants	1q21.1	NBPF9	ASD, ID, SCZ, ADHD, EPI
	2p16.3	NRXN1	ASD, ID, language delay, SCZ.
	3p13	FOXP1	ID, ASD, SLI
	6q16.3	GRIK2	Recessive ID
	7q11.23	FKBP6/CLIP2	ASD, ID, language delay
	7q31.1	FOXP2	SLI
	11q13.3-q13.4	SHANK2	ASD, ID
	15q11-q13	MAGEL2/ NDN	ASD, EPI, ID
	16p11.2	VPS35/ORC6	ASD, ADHD, ID, EPI, SCZ
	16p13.3	A2BP1	ID, ASD, EPI, SCZ, ADHD
	17q11.2	SLC6A4	ASD, OCD
	17q12	ACCN1/PNMT	ASD, SCZ, EPI
	22q11.21		DiGeorge syndrome, SCZ, ASD, ID.BPAD
	22q13.33	SHANK3	ASD, Phelan McDermid syndrome**
	Xq13.1	NLGN3	ASD
	Xp22.11	PTCHD1	ASD, ID
	Xp22.32-p22.31	NLGN4X	ASD, ID, TS, ADHD
Common Alleles	1q42.2	DISC1	SCZ,BPAD
	2q31.1	SLC25A12	ASD
	3p25.3	OXTR	ASD
	7q31.2	MET	ASD, Diabetes II
	7q22.1	RELN	ASD
	7q36.3	EN2	ASD
	12q14.2	AVPR1A	ASD
	17q21.32	ITGB3	ASD

Abbreviations: LTD, long-term depression; LTP, long-term potentiation; PPI, prepulse inhibition; E/I, excitatory/inhibitory; PSD, postsynaptic density; ASD, autism spectrum disorders; SCZ, schizophrenia; ADHD, attention deficit hyperactivity disorder; ID, intellectual disability; XLID, X-linked intellectual disability; LIS, lissencephaly; EPI, epilepsy; OCD, obsessive compulsive disorder; TS, Tourette syndrome; SLI, speech and language impairment; USV, ultrasonic vocalization; TF, transcription factor; ECM, extracellular matrix; GPCR, G-protein-coupled receptor;BPAD, Bipolar affective disorder.

**A rare autosomal dominant inherited disorder characterized by multiple tumor-like growths, increased risk of certain forms of cancer, and diverse clinical features including neurologic features such as autism and Lhermitte Duclos disease* (Tsuchiya et al., 1998 & Zhou et al.,2001).

*** A genetic syndrome caused by disruption of the SHANK3 gene which codes for the shank3 protein. The protein most important role is in the brain. It is involved in processes crucial for learning and memory. It also has an important role in brain development. It is also known as 22q13.3 deletion syndrome and is highly associated with autism.*

Human (Homo sapiens) Genome Browser Gateway, http://genome.ucsc.edu/cgi-bin/hgGateway.

Table B1. The five segmental duplication (SD)-rich regions used in this study that are implicated in risk for ASD.

Chromosome #	Start	End	Length (bp)
7	61058424	81999980	20,941,556
10	77000071	91999901	14,999,830
15	18260026	34999924	16,739,898
17	12000112	22187009	10,186,897
22	14430001	25999992	11,569,991

**bp: base-pairs*.