Wavelets Based Feature Extraction With PCA For Predicting Autism In Neonates Using Navïe Bayes Classifier

Isabel Mensah
National Institute for Mathematical Sciences
Peter Amoako-Yirenkyi (amoakoyirenkyi@knust.edu.gh)
National Institute for Mathematical Sciences
Nana K Frempong
National Institute for Mathematical Sciences
George P Lamptey
National Institute for Mathematical Sciences

Research Article

Keywords: Autism Spectrum Disorders, Wavelet Transform, Principal Component Analysis, Naive Bayes classifier, Classification, K-means clustering

Posted Date: November 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1048775/v1

RESEARCH

Wavelets based feature extraction with PCA for predicting autism in neonates using Navïe bayes classifier

Isabel Mensah[1,5†], Peter Amoako-Yirenkyi[1,2,4*], Nana K Frempong[1,3] and George P Lamptey[1,2]

Abstract

Background: Current studies show early interventions of autism increase significant long-term positive effects, symptoms and, later skills. Currently, These interventions are based on the use of an early diagnostic test. Existing methods for diagnosing Autism Spectrum Disorders (ASDs) such as cognitive tests, Intelligence Quotient, and standardized tests like the Autism Diagnostic Observation Schedule (ADOS) are functionally limited since they rely on child development for diagnoses. The standard is that a child must be at least three(3) years to undergo these tests. Accurate diagnosis is only possible after this period, and this may contribute to delayed diagnosis with an overall effect on the health system. In this era of increasing genetic data, it is possible to infer the genetic patterns of the disorder. This study introduces a novel and rigorous approach for predicting ASDs in neonates and their subsequent severity by identifying significant genes that contribute to the disorder.

Methods: We used a wavelet transform and t-test to identify the significant genes that contribute to the disease. We subsequently employed the Naive Bayes classifier in the prediction of the autistic status of the neonate.

Additionally, Principal Component Analysis (PCA) was employed to remove all the dependencies among the genes to enhance classification. Finally, we used the K-means clustering method to predict the severity level of the disease in the neonate.

Results: Up to 200 differentially expressed genes were identified and used for predicting the ASD status of the child with a classification accuracy of 95.91%. Also, the results of the K-means demonstrated that the higher the mean of the cluster, the more severe the disease would be among that corresponding group. Optimizing and implementing these models in clinical settings may significantly reduce the health burden of ASDs.

Keywords: Autism Spectrum Disorders; Wavelet Transform; Principal Component Analysis; Naive Bayes classifier; Classification; K-means clustering

Introduction

Autism Spectrum Disorders (ASDs), as first explained by Kanner [18], is pervasive developmental disabilities characterized by the awkwardness of social relations and communication skills, restricted recurrence interest, and behavior. It is a family of complex disorders of brain development.

In reference to the Diagnosis and Statistical Manual of developmental disorders (DSM)-IV-TR criteria, pervasive disorders have been grouped into five different disorders with their respective diagnoses. They include; Rett syndrome, Autistic disorder (classic autism), pervasive developmental disorder (PDD-NOS), Asperger syndrome, and childhood disintegrative syndrome.

Complications in motor coordination and attention, physical health problems like sleep and gastrointestinal disorders, and intellectual disabilities can be related to ASD. The causes of ASD have remained somewhat a mystery. However, it has been known to be caused by some environmental and genetic factors. Such factors include immune dysfunction, prenatal and perinatal factors, drugs and toxic exposure at pregnancy, very low birth weight, and advanced maternal age.

*Correspondence: amoakoyirenkyi@nims.edu.gh
[1]National Institute for Mathematical Sciences, Ghana, Kumasi, Ghana
Full list of author information is available at the end of the article
†Equal contributor

ASD is a heterogeneous condition that affects 1 in 68 children globally [3]. Unlike other disorders among children, such as blindness and malformations, which are detected at birth, ASD detection is based on some overt characteristics that occur later in the child's life. Additionally, many children with ASD might not get the help they need if the diagnosis is delayed. The earlier an ASD is diagnosed, the sooner treatment can begin. Indeed, the American Academy of Pediatrics (AAP) recommends that all children should be screened for developmental delays and disabilities during their regular well-child doctor visits at 9, 18, 24/30 months [23].

Even though ASD diagnosis can be made at all ages, it is not reliable, particularly for children below 18 months. This is because doctors basically diagnose the disease by looking at a person's behavior and development. ASDs diagnosis and its associated severity are mainly made by cognitive tests and standardized tests such as the Autism Diagnostic Observation Schedule (ADOS) [22]. Because of this, other experimental and computational methods other than the approved screening instruments can be put in place to diagnose the disease more reliably at an earlier stage for treatment. Genetically, research has shown that more than two genes cause the disorder, and these genes interact in a complex manner. Several genes, between two and hundreds, have been identified and could contribute to disease susceptibility. Family studies have it that genetic disorders are strongly associated with ASDs, and the occurrence risk in siblings with an ASD patient proband between 8-10% [33]. A more recent study has observed up to 25% of siblings have been affected [9]. In identifying gene-interrelated diseases like ASD, biologists pay attention to gene expression data relating to the diseases. They usually use some time-consuming and expensive methods such as copy number variation studies (CNV), whole-exome sequencing (WES), and genome-wide association studies (GWAS). However, computational techniques are faster, reliable, and provide inexpensive solutions for predicting potential candidate disease genes. Ansel et al. [2] conducted an extensive review on Transcriptomics Studies for the variability of Gene Expression in ASD. They explained ASDs as diversified and emerge from epigenetic, genetic, and environmental origins. Yet, as stated earlier, the exact causes of ASDs remain elusive. The assessment of an individual's behavior and phenotype has been the basis for diagnosing ASD, as elaborated in their work. They found numerous ASD susceptibility genes and for the majority of the ASD population. The genes identified are known to be involved in a broad and diverse range of biological functions.

Burgeoning research has focused on classifying these genes with various mathematical methods into diseased and non-diseased genes, which aids in predicting the presence of a disease in an individual. Nanni and Lumini [27], studied disease classification by DNA micro-array data using the wavelet transform selection method. It was noted that the huge dimension of the feature vector mostly contained some information that was not relevant for accurate classification. Hence, they employed the wavelet transform to select the relevant features for the classification. Likewise, Bennet et al. [6] also used a discrete wavelet technique for feature extraction and then employed a hybrid classifier for micro-array data analysis for Cancer. The method which was proposed in their work was based on the naive Bayes, support vector machine (SVM), and K nearest neighbor (KNN). They combined the discrete wavelet transform (DWT) and a Moving Window Technique (MWT) for feature selection. Much in the same way, Hameed et al. [12] analyzed the genes expressed in ASD to select the most important genes for classification. They achieved their objective using several statistical filters and a geometric binary particle swarm optimization-support vector machine (GBPSO-SVM) algorithm. Gok [11] tried to predict the presence of some disease risk genes that contribute to ASD. In his work, he trained a model with brain developmental gene expression data for classifying ASD risk genes using some machine learning techniques. His results confirmed the model's performance with 0.902 sensitivity, 0.839 area under Receiver Operating Characteristic (ROC) curve, Matthews correlation coefficient (MCC) of 0.583, and an F-measure score of 0.806 of the presence of the disease in the individual.

This paper presents a wavelet-based feature extraction method with principal component analysis to predict the possibility of ASD in neonates using the naive Bayes classifier and its associated severity.

Results and Discussion

This section presents the analysis and results for some autistic and healthy individuals used in the classification problem. In addition, the methods discussed in the methodology sections were implemented. Finally, we discuss the analysis and interpretation of the results to conclude the section.

Feature extraction and Selection

It is known that the gene expression dataset is enormous, noisy, and has a lot of irrelevant information, which potentially compromises the classification process in terms of accuracy. Hence, preprocessing is done to extract the essential information relevant to the study. This will nonetheless also reduce the noise and

variations in the dataset.
Statistical methods for exploratory analysis of multidimensional data work best for data with the same range of variance at different ranges of the mean values. The amount of variance is expected to be approximately the same across different mean values. Thus, the data must be homoskedastic, and hence log 2 is applied to transform/normalize the data.
Upon transforming the data, the Haar wavelet transform was used to compute the wavelets coefficients as shown in Figure 1;

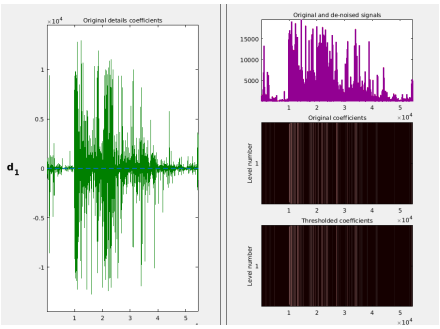

Figure 1: Wavelet coefficient before and after thresholding

A threshold value, λ, thus the standard deviation for the wavelet coefficient was set to remove the coefficient that are less than the threshold. It is assumed that, all the wavelet coefficients below the threshold are regarded as noise and not depicting the true expression level of the genes.
After thresholding, a new set of wavelet and approximate coefficients are obtained and the inverse wavelet transform is taken with these coefficients to get an approximate signals or dataset retaining its dimension.

In summary, after the transformation and extraction, it was realized that, on average, the genes became closer to the means in each sample. Hence, the feature selection method was applied to select the significant genes for the study using the t-test. Thus, the T-test was used to assess if a particular gene is significantly different in the two classes, the autistic and non-autistic classes. The independent t-test selected 1,973 out of the 54,613 genes in the 146 samples as the significant genes whose means differ in the two-class cases. The 1,973 selected genes were ranked according to their P-value. It was assumed that the smaller the P-value, the more significant the gene is. This ranged between $5.8e-06$ and $3.4e-02$. From this range, the first 200 most significant genes were selected as they were much relevant to the study.

Classification on T-test with and without ranking

The classification was made on the dataset selected by t-test with a resultant dimension of 1,974 genes and 146 samples using the Navïe Bayes classifier (classification without ranking). This set of features were used in the model for classification. The dataset was divided into 60% training and 40% testing datasets, with 97 and 49 samples, respectively. The performance of the model has been explained in table 1.
From table 1, the classifier was able to predict the presence of the disease in a sample with accuracy and precision of 63.27 and 65.21 respectively on the 1,973 genes. However, the significant genes tested were tried on the classifier to improve its accuracy.

Table 1: Performance of the classifier with the t-test results

	Classification without Ranking	
	Training set	Testing set
Accuracy	83.50%	63.27%
Sensitivity	84.60%	60.00%
Specificity	82.20%	66.67%
Precision	84.61%	65.21%

Again, as discussed, the 200 selected genes were inputted into the model for classification with 60% training and 40% testing comprising 97 and 49 samples, respectively. The performance of the model has also been explained in the table 2. It can be seen that the classifier increased its accuracy and precision by ten percentage points, thus, 73.47% and 77.27% respectively on the testing data.

Table 2: Performance of the classifier with 200 significant genes

	Classification without Ranking	
	Training set	Testing set
Accuracy	86.59%	73.47%
Sensitivity	86.50%	68.00%
Specificity	82.98%	79.17%
Precision	88.24%	77.27%

Classification with PCA

Dependencies were tested and it was noted that, there were some moderate and very weak correlation among the genes which may have contributed to the average performance of the classifier. Hence, the principal component analysis was computed on the dataset to eliminate all dependencies in the data.
The PCA identified 114 component explaining 95.4% of the variations in the data and the performance of the model on the set has been shown in table 3 .

Table 3: Performance of the classifier on PCA results

	Classification without Ranking	
	Training set	Testing Set
Accuracy	98.69%	95.91%
Sensitivity	98.10%	96.00%
Specificity	100.00 %	91.67%
Precision	100.00 %	92.6 %

In summary, it has been revealed that performance of the classifier increased drastically when all the dependencies in the dataset was removed by PCA.

K-Means Clustering

After the samples has been correctly identified, a further step is taken if the person is an autistic patient. This involves another level of classification to determine whether the person's autism will be mild, moderate or severe. The k-means cluster was adopted due to the absence of a prior knowledge on the labels in the dataset.

Choosing the K

The K-means clustering algorithm as discussed earlier was used to classify or group the objects based on features that are partitioned into K number of groups where K is a positive integer. Since we want to group the autistic class into the patient's severity as mild, moderate and severe, K was set to 3 and total number of 77 autistic samples were used.
After the clustering, cluster 1, 2 and 3 ((cluster mild, moderate and severe) got 18, 39 and 20 samples respectively. Example on how the sample samples where distributed among the clusters has been shown in the clustering vector shown in table 5. Its records the individual autistic patients and the cluster they belong.

Table 4: Clustering Vector

sample number	cluster
70	2
71	2
72	2
73	2
101	3
102	1
103	3
104	1

The assumption made on the K-Means Clustering was that, the cluster with the highest mean will be the group that is highly autistic (severe) and the vice versa.

Table 5: Cluster groups and associated severity

cluster	size	mean	severity
C1	18	5.598028	mild
C2	39	5.627063	moderate
C3	20	5.642697	severe

Conclusions

Early recognition and diagnosis of ASD in children are necessary for expeditious behavioral and educational interventions, including referral to a formal early intervention program, with the potential resultant improving the prognosis, especially for cognition, peer interactions, and language development. Knowing that a child has a specific diagnosis and is receiving therapy can also help a family cope better. Without knowing that the child's diagnosis increases parental anxiety and delays, the introduction of interventions can reduce behavioral problems and optimize outcomes in the child. In addition, literature has it that having a child with autism in a family increases the risk of other siblings having a disorder on the autism spectrum or with the broader phenotype. Early identification could prompt parents to receive genetic counseling and plan for future children.

ASD has a genetic trait, and it is crucial to understand the genetic biomarkers underlying the disease. Information on the genetic biomarkers can serve as a reasonable basis for predicting ASD at a relatively cheaper cost. This study analyzes efficiently and predicts the presence of ASD in an individual utilizing these biological markers with mathematical models and the severity of the disease. As part of the analysis, the feature extraction and selection methods employed reduced the number of genes from 54,613 to 1,973 significant genes. The Navïe Bayes classifier was adopted and was able to classify the presence or absence of the disease accordingly in an individual with an accuracy of 95.91%. The severity of the presence of ASD was determined alongside its diagnosis utilizing the K-means clustering algorithm. This prediction indicated that samples clustered around the highest mean are more likely to represent high autistic status and vice versa.

Further analysis and testing have to be conducted on the severity prediction to establish the assumption made on the test. In particular, a dataset with labels on the patient's severity level should be collected to confirm the assumption made on the clustering. The dataset from neonates can also be collected to test the model for further studies. The prediction model is open for further interrogation and should be tried in clinical settings to test the effectiveness of identifying the significant genes. It will help provide a deeper understanding of the abnormal expression patterns of

the disease, paving the way for drug discovery where special autism drugs can be designed to control and correct abnormal gene expression. The study also recommends that early detection of ASD at birth using algorithms should be given due consideration to complement other existing methods.

Methods

In this section, the proposed model for diagnosing the disease has been described and the gene expression-Autism dataset used.

Proposed Model

The proposed model, as illustrated in Figure 2 consist of six main steps: (1) extraction of features into one-dimensional Haar Wavelet Transforms (HWT) feature vector and obtaining a set of wavelet and scaling coefficients; (2) Reconstruction of data using the new set of wavelet and scaling coefficients; (3) The selection of differentially expressed genes that collectively contribute to the disease using the independent t-test; (4) Removing dependencies from the features using PCA; (5) Training the naive Bayes classifier to make predictions; (6) determining the possible severity of the disease in the neonate using the k-means clustering. In Figure 2, the preprocessing stage of the model is composed of HWT, inverse HWT, t-test, and PCA to basically extract and select the relevant features and also to reduce noise in the dataset to improve on the accuracy of the classifier. Only a few studies combine some preprocessing methods with a classifier to build ASD risk gene prediction. Still, this work improves on the preprocessing done in previous studies, which makes it more robust for predicting the presence of ASD in neonates and its possible severity of the disease.

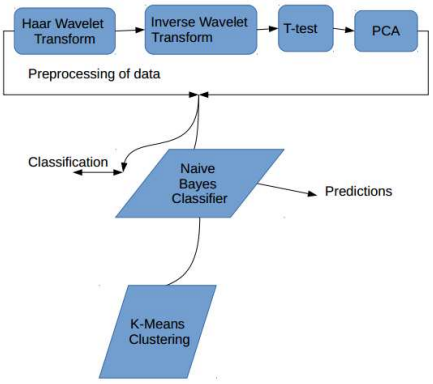

Figure 2: Proposed Model

Experimental data

The experimental data used in the study is made up of an autism micro-array dataset obtained from the Gene Expression Omnibus hosted by NCBI [1]. The samples are people in the Phoenix area of the south-western U.S. state of Arizona. The blood sample from the observations was collected in the spring and summer of 2004. RNA was totally extracted for the micro-array experiment using Affymetrix Human U133 Plus 2.0 39 Expression Arrays. The dataset is made up of 146 samples (observations) with 54,613 genes. The observations are carved up into two classes(the control class and the autistic class). According to the DSM-IV criteria, the autistic patients taken were diagnosed by medical practitioners and were confirmed based on the ADOS and ADI-R criteria.

Haar Wavelet Transforms

To provide the reader opportunity to understand the full scope of the work, we briefly discuss the Haar Wavelet Transform(HWT) and associated functions for constructing it. HWT is a method that transforms a digital signal into a vector space and ensures that the high-frequency and low-frequency components are separated. HWT, which is discrete in nature, is applied to the data to find the most discriminant features between the two classes. HWT makes scaling or translation to the signals to obtain their orthonormal basis representation using the Haar wavelet function $w(t)$ over an interval.

The orthogonal set of Haar functions are defined in the interval $x \in [0,1]$ For every pair of $j,k \in Z$, the Haar function $\psi_j, k(x)$ is defined as

$$\psi_j, k(x) \quad = \quad 2^{j/2}\psi(2^j(x)-k),\, t \in R \tag{1}$$

The function is supported on the right open interval;

$$I_n \quad = \quad \left[k2^{-j}, (k+1)2^{-n}\right) \tag{2}$$

The family of $\psi_{j,k}(t)$, constitutes an orthonormal basis of $L^2(R)$ such that.

$$\int \psi_{j,k}(t)\psi^*_{m,n}(t) = \begin{cases} 1; & j=m, k=n \\ 0; & otherwise \end{cases} \tag{3}$$

The scaling function is defined as;

$$\psi(t) = \begin{cases} 1, & t \in [0,1) \\ 0, & otherwise \end{cases} \tag{4}$$

$$\tag{5}$$

And the Haar mother wavelet is given as;

$$\psi(t) = \begin{cases} 1, & x \in [a, b) \\ -1, & x \in [b, c) \\ 0, & otherwise \end{cases} \tag{6}$$

Haar wavelet transform extracted as a matrix in equation (3) serves as the wavelet filter used in the convolution to produce the coefficients. A universal threshold technique of the standard deviation is applied to filter the expression values which are not significant to the study. In so doing, the important coefficient or information is extracted from the dataset. The inverse wavelet transform is applied to reconstruct the signal from the extracted coefficient. After the signal has been reconstructed, a feature selection technique is applied to reduce the dimension.

T - test and Hypothesis Testing

- H_0: There is no significant difference between the means of the autistic class and non-autistic class for $gene_i$
- H_A: There is a significant difference between the means of the autistic class and non-autistic class for $gene_i$

Test Statistic
Two- sample independent t-test for $gene_g$ for autistic class(i) and non-autistic class(j) is presented in equation (7)

$$t_g = \frac{\overline{X_{ig}} - \overline{X_{jg}}}{\sqrt{\frac{s_{ig}^2}{n_i} + \frac{s_{jg}^2}{n_j}}} \tag{7}$$

where;

- $\overline{X_{ig}}, \overline{X_{jg}}$ - means of autistic class(i) and non-autistic class(j) respectively for a given gene
- s_{ig}^2, s_{jg}^2 - Standard deviation of autistic class(i) and non-autistic class(j) respectively for a given gene
- n_i, n_j - number of samples of autistic class(i) and non-autistic class(j) respectively for a given gene
- $\frac{s_{ig}^2}{n_i}, \frac{s_{jg}^2}{n_j}$ - Standard error terms SE_i and SE_j

Navïe Bayes Classifier

The Navïe Bayes classifier is a simple probabilistic classifier based on applying the Bayes theorem where every feature is assumed to be class-conditionally independent.

The classifier employs the posterior probabilities to assign the class label to a test pattern; a pattern is assigned the class label with the maximum posterior probabilities. The Posterior Probability of a person being Autistic has been set out in the equation 8.

$$p(A \mid f) \propto p(f \mid A)P(A) \tag{8}$$
$$p(A^c \mid f) \propto p(f \mid A^c)P(A^c) \tag{9}$$

Where;
A = Autistic class
A^c = Non- Autistic class
f = feature set of Genes

Given that the feature (a set of genes) is a vector $f = (G_1, \cdots G_g)$, the criterion for classifying whether or not a person is Autistic is explained as;

$$p(C_j) = argmax_j \prod_{i=1}^{g} p(G_i \mid A_j)p(A_j) \tag{10}$$

where
$j = 1, 2$
$A_1 = A$, $A_2 = A^c$

Principal Component Analysis (PCA)

PCA is employed to take away all dependencies in the dataset to enhance the performance of the classifier. The PCA algorithm has been explained below;

- The selected features after the application of the T- test is used.
- Find the empirical mean of each gene $j = 1, \cdots, n$ and calculate the deviation from the mean.
- Calculate the covariance and find the covariance matrix of the $n \times p$ matrix.
- Calculate the eigenvalues and eigenvectors of the covariance matrix.
- Choose the components and form a feature vector.
- Derive a new data set with the feature vector.

K-Means Clustering

In order to predict the level of severity among the Autistic patients, the K-means clustering algorithm was used since there were no initial labels that came with the dataset. The K-means Process has been explained in figure 3

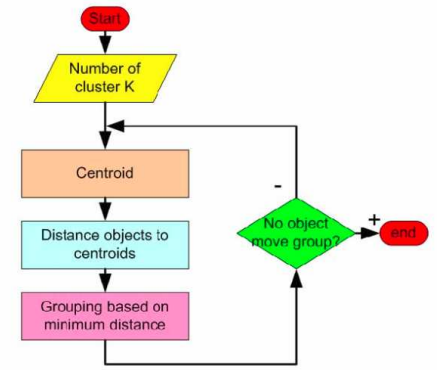

Figure 3: K-means process

Accessing the Performance of the Classifier

The confusion matrix shown in table (6) was used in evaluating the performance of the classification model on a set of test data for which the true values are known.

Table 6: Confusion Matrix

	ACTUAL VALUES	
PREDICTED VALUES		
	POSITIVE	NEGATIVE
POSITIVE	True positive	False positives
NEGATIVE	False negative	True Negative

- True Positives: These are samples in which the model predicted to be autistic and were truly autistic.
- True Negatives: The samples the model predicted as non autistic and were truly non autistic.
- False positive: The samples predicted as autistic but were non autistic. (Type I error).
- False Negative: The samples predicted as non autistic but were Autistic (Type II error).

Performance Measures

- The Sensitivity (True positive rate): It measures the proportion of autistic samples that have been identified as autistic.

$$Sensitivity = \frac{TruePositive}{(TruePositive + FalseNegative)} \quad (11)$$

- The Specificity (True negative rate): It measures the proportion of percentage of the non autistic sample that have been identified as non autistic.

$$Specificity = \frac{TrueNegative}{(TrueNegative + FalsePositive)} \quad (12)$$

- Precision: It captures the proportion of the positive predictive values of the positive predictions that are actually positive

$$Precision = \frac{TruePositive}{(Truepositive + FalsePositive)} \quad (13)$$

- Classification accuracy: It describes how often the classifier predicts correctly.

$$Accuracy = \frac{TruePositive + Truenegative}{Totalsamples} \quad (14)$$

Acknowledgements
We will like to acknowledge the National Institute of Mathematical Sciences, Ghana for providing computing resources for this study.

Funding
Not applicable.

Abbreviations
- ASD: Autism spectrum Disorders
- PCA: Principal Component Analysis
- DWT: Discrete Wavelet Transform
- ADOS: Autism Diagnostic Observation Schedule
- PDD-NOS: Pervasive Developmental Disorder

Availability of data and materials
The micro-array gene expression data used and analyzed during the current study is available at the Gene Expression Ominibus (GEO) under the accession number GSE25507. [Link to the data]

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that the authors have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Consent for publication
Not applicable

Authors' contributions
All authors were involved in Conceptualization, Methodology, Investigation, Formal analysis, Original draft preparation, and reviewing of the manuscript.

Author details
[1]National Institute for Mathematical Sciences, Ghana, Kumasi, Ghana. [2]Department of Mathematics, KNUST, Kumasi, Ghana. [3]Department of Statistics and actuarial Science, KNUST, Kumasi, Ghana. [4]Center for Scientific Computing and Industrial modeling, NIMS, GHANA, Kumasi, Ghana. [5]Department of Biochemistry and Biotechnology, KNUST, Kumasi, Ghana.

References

1. Alter, M.D., Kharkar, R., Ramsey, K.E., Craig, D.W., Melmed, R.D., Grebe, T.A., Bay, R.C., Ober-Reynolds, S., Kirwan, J., Jones, J.J., *et al.*: Autism and increased paternal age related changes in global levels of gene expression regulation. PloS one **6**(2), 16715 (2011)
2. Ansel, A., Rosenzweig, J.P., Zisman, P.D., Melamed, M., Gesundheit, B.: Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies. Frontiers in neuroscience **10**, 601 (2017)
3. Baio, J.: Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010 (2014)
4. Benaron, L.D.: Autism (Biographies of Disease), (2008)
5. Benedetto, J.J.: Wavelets: Mathematics and Applications vol. 13. CRC press, ??? (1993)
6. Bennet, J., Arul Ganaprakasam, C., Arputharaj, K.: A discrete wavelet based feature extraction and hybrid classification technique for microarray data analysis. The Scientific world journal **2014** (2014)
7. Cichosz, P.: Data Mining Algorithms: Explained Using R. John Wiley & Sons, ??? (2014)
8. Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection. IEEE Transactions on information theory **38**(2), 713–718 (1992)
9. Constantino JN, F.T.A.A.L.P. Zhang Y: Sibling recurrence and the genetic epidemiology of autism. Am J Psychiatry **1349-56**, 167–11 (2010)
10. Dr Ananya Mandal, M.: Autism history. Technical report, New medical life sciences (Febauary 2018)
11. Gok, M.: A novel machine learning model to predict autism spectrum disorders risk gene. Neural Computing and Applications (2018)
12. Hameed, S.S., Hassan, R., Muhammad, F.F.: Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a gbpso-svm algorithm. PloS one **12**(11), 0187371 (2017)
13. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, ??? (2011)
14. Hu, V.W., Sarachana, T., Kim, K.S., Nguyen, A., Kulkarni, S., Steinberg, M.E., Luu, T., Lai, Y., Lee, N.H.: Gene expression profiling differentiates autism case–controls and phenotypic variants of autism spectrum disorders: Evidence for circadian rhythm dysfunction in severe autism. Autism research **2**(2), 78–97 (2009)
15. Ingrid, D.: Ten Lectures on Wavelets, 1st edn. CBMS-NSF regional conference series in applied mathematics 61. Society for Industrial and Applied Mathematics, ??? (1992)
16. Jackson, J.E.: Principal components and factor analysis: part iprincipal components. Journal of Quality Technology **12**(4), 201–213 (1980)
17. Jose, A.: Gene selection by 1-d discrete wavelet transform for classifying cancer samples using dna microarray date. PhD thesis, University of Akron (2009)
18. Kanner, L., *et al.*: Autistic disturbances of affective contact. Nervous child **2**(3), 217–250 (1943)
19. Kim, T.K.: T test as a parametric statistic. Korean journal of anesthesiology **68**(6), 540–546 (2015)
20. Kumar, G., Bhatia, P.K.: A detailed review of feature extraction in image processing systems. In: Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference On, pp. 5–12 (2014). IEEE
21. Lokenath Debnath, F.A.S.: Wavelet Transforms and Their Applications Second Edition. Springer, ??? (2015)
22. Lord, C., Risi, S.: The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders **30**(3), 2 (2000)
23. Lord, C., Risi, S., DiLavore, P.S., Shulman, C., Thurm, A., Pickles, A.: Autism from 2 to 9 years of age. Archives of general psychiatry **63**(6), 694–701 (2006)
24. Miranda, A.A., Le Borgne, Y.-A., Bontempi, G.: New routes from minimal approximation error to principal components. Neural Processing Letters **27**(3), 197–207 (2008)
25. Montgomery, D.C., Runger, G.C.: Applied Statistics and Probability for Engineers. John Wiley & Sons, ??? (2010)
26. Murty, M.N., Devi, V.S.: Pattern Recognition: An Algorithmic Approach. Springer, ??? (2011)
27. Nanni, L., Lumini, A.: Wavelet selection for disease classification by dna microarray data. Expert systems with applications **38**(1), 990–995 (2011)
28. [online] Available at: http://autismspeaks.org, H..A.S.: Autism speaks (2015)
29. Pandey, A.K., Pandey, P., Jaiswal, K.: Classification model for the heart disease diagnosis. Global Journal of Medical Research (2014)
30. Plauche Johnson, C.: Early Clinical Characteristics of Children with Autism (2004). 83-121
31. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science **270**(5235), 467–470 (1995)
32. Shen, M., Piven, J.: Brain and behavior development in autism from birth through infancy **19**, 325–333 (2017)
33. SzatmariM P, Z.L.M.J. Jones MB: Genetics of autism: Overview and new directions. Journal of Autism and Developmental Disorders **28** (1998)
34. Yim, K.H., Nahm, F.S., Han, K.A., Park, S.Y.: Analysis of statistical methods and errors in the articles published in the korean journal of pain. The Korean journal of pain **23**(1), 35–41 (2010)