GENE SELECTION AND CLASSIFICATION IN AUTISM GENE EXPRESSION DATA

SHILAN SAMEEN HAMEED AL-JAF

A dissertation submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Computer Science

Faculty of Computing
Universiti Teknologi Malaysia

OCTOBER 2017

ACKNOWLEDGEMENTS

"In the name of Allah, the most Gracious and the most Merciful"

First and foremost, I must thank the almighty Allah for blessing me with His continuous help to achieve my goals. Special thank goes to my supervisor Dr. Rohayanti Hassan for her splendid support and encouragement during my research works. This study would not have been completed without her excellent supervision, belief, patience and confidence to me.

I would also like to express my deep gratitude to my husband Fahmi F. Muhammad for his endless love, patience and support to overcome the difficulties and challenges throughout my study period.

I should never forget to thank my little daughter and my mother for their sacrifices and understanding in standing with me to reach this stage.

I am also grateful to all staffs at the Faculty of Computing (FC), the Universiti Teknologi Malaysia (UTM) for their kind cooperation during this special journey.

ABSTRACT

Autism spectrum disorders (ASD) are neurodevelopmental disorders that are currently diagnosed on the basis of abnormal stereotyped behaviour as well as observable deficits in communication and social functioning. Although a variety of candidate genes have been attributed to the disorder, no single gene is applicable to more than 1–2% of the general ASD population. Despite extensive efforts, definitive genes that contribute to autism susceptibility have yet to be identified. The major problems in dealing with the gene expression dataset of autism include the presence of limited number of samples and large noises due to errors of experimental measurements and natural variation. In this study, a systematic combination of three important filters, namely t-test (TT), Wilcoxon Rank Sum (WRS) and Feature Correlation (COR) are applied along with efficient wrapper algorithm based on geometric binary particle swarm optimization-support vector machine (GBPSO-SVM), aiming at selecting and classifying the most attributed genes of autism. A new approach based on the criterion of median ratio, mean ratio and variance deviations is also applied to reduce the initial dataset prior to its involvement. Results showed that the most discriminative genes that were identified in the first and last selection steps concluded the presence of a repetitive gene (CAPS2), which was assigned as the most ASD risk gene. The fused result of genes subset that were selected by the GBPSO-SVM algorithm increased the classification accuracy to about 92.10%, which is higher than those reported in literature for the same autism dataset. Noticeably, the application of ensemble using random forest (RF) showed better performance compared to that of previous studies. However, the ensemble approach based on the employment of SVM as an integrator of the fused genes from the output branches of GBPSO-SVM outperformed the RF integrator. The overall improvement was ascribed to the selection strategies that were taken to reduce the dataset and the utilization of efficient wrapper based GBPSO-SVM algorithm.

ABSTRAK

Gangguan spektrum autisme (ASD) adalah gangguan perkembangan neuro yang kini didiagnos berdasarkan tingkah laku stereotaip yang tidak normal serta dilihat sebagai kelemahan dalam aspek komunikasi dan fungsi sosial. Walaupun telah banyak kajian terhadap pelbagai gen calon telah dikaitkan dengan kelemahan ini, tiada gen tunggal yang boleh digunakan mewakili lebih dari 1-2% populasi ASD. Walaupun banyak usaha telah dilaksanakn, gen definitif yang menyumbang kepada kecenderungan autisme belum dikenalpasti. Permasalahan utama dalam menangani dataset ekspresi gen autisme termasuklah kehadiran jumlah sampel yang terhad dan noisy data akibat ralat dalam pengukuran dan variasi semulajadi. Dalam kajian ini, satu gabungan sistematik tiga saringan penting iaitu t-test (TT), Wilcoxon Rank Sum (WRS) dan Ciri Korelasi (COR) digunakan bersama dengan algoritma pembalut efisien yang berdasarkan mesin vektor sokongan dan dioptimum dengan zarah binari geometrik (GBPSO-SVM), bertujuan untuk memilih dan mengklasifikasikan gen autisme yang paling signifikan. Pendekatan baru berdasarkan kriteria nisbah median, nisbah min dan penyimpangan varians juga digunakan untuk mengecilkan amaun dataset. Keputusan menunjukkan bahawa gen paling diskriminatif yang dikenalpasti dalam langkah pemilihan pertama dan terakhir mendapati kehadiran gen berulang (CAPS2), iaitu sebagai gen ASD paling risiko. Hasil gabungan daripada subset gen yang dipilih oleh algoritma GBPSO-SVM didapati telah meningkatkan ketepatan klasifikasi ke sekitar 92.10%, iaitu lebih tinggi daripada yang dilaporkan dalam kajian literatur terdahulu untuk dataset autisme yang sama. Kenyataannya, penerapan metod himpunan menggunakan algoritma Random Forest (RF) menunjukkan prestasi yang lebih baik berbanding dengan kajian terdahulu. Walau bagaimanapun, prestasi metod himpunan menggunakan SVM sebagai penyepadu gen yang bersatu daripada cawangan pengeluaran GBPSO-SVM mengatasi penyepadu RF. Secara keseluruhannya, penambahbaikan telah berjaya dilaksana sebagai strategi baru dalam pemilihan untuk mengurangkan dataset dan penggunaan algoritma GBPSO-SVM berdasarkan pembalut yang efisien.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

a.k.a.	-	Also Known As
AUC	-	Area Under Curve
ADI-R	-	Autism Diagnostic Interview, Revised
ADOS	-	Autism Diagnostic Observation Schedule
ASD	-	Autism Spectrum Disorder
BPSO	-	Binary Particle Swarm Optimization
cDNA	-	complementary Deoxyribonucleic Acid
CFS	-	Correlation based Feature Selection
CC	-	Correlation Coefficient
DNA	-	Deoxyribonucleic Acid
DSM-IV	-	Diagnostic and Statistical Manual of Mental Disorders fourth version
FN	-	False Negative
FN	-	False Positive
FCBF	-	Fast Correlation Based Filter
COR	-	Feature Correlation with Class
FS	-	Feature Selection
FDA	-	Fisher Discriminant Analysis
GEO	-	Gene Expression Omnibus
GA	-	Genetic Algorithm
GBPSO	-	Geometric Binary Particle Swarm Optimization
gbest	-	Global Best
KNN	-	K-nearest Neighbour
mRNA	-	Messenger Ribonucleic Acid
MA	-	Microarray
NCBI	-	National Centre for Biotechnology Information
NP	-	Non-deterministic Polynomial
OOB	-	Out of Bag

PSO	-	Particle Swarm Optimization
pbest	-	Personal Best
PDD	-	Pervasive Developmental Disorder
RF	-	Random Forest
ROC	-	Receiver Operating Characteristics
RFE	-	Recursive Feature Elimination
RNA	-	Ribonucleic acid
SBS	-	Sequential Backward Selection
SFS	-	Sequential Forward Selection
SVM	-	Support Vector Machines
TN	-	True Negative
TP	-	True Positive
TT		Two Sample T-test

CHAPTER 1

INTRODUCTION

1.1 Background

Autism is a neuro-developmental disorder which is defined by a weakened social interaction, impaired verbal and non-verbal communications as well as repetitive actions among the autistic persons. It is also recognized as autistic spectrum disorder (ASD) or pervasive developmental disorder (PDD) (Muhle *et al.*, 2004). This disorder is appeared in more than 1% of the population, whereas males are four times more vulnerable to the disorder compared to females (De Rubeis and Buxbaum, 2015). It has been reported that the prevalence rate of autism has dramatically risen in the last decade from the year of 2000 to 2012, as shown in Figure 1.1. Generally, the symptoms of autism disorder will be apparently seen in early age of childhood, especially before age three, at which the diagnosing purpose can be initiated. A wide range of phenotypes and intellectual disability (of about 35%) is persistent with the autistic people, while their language delay is counted to be 50% and epilepsy is from 5–15% (Geschwind and State, 2015).

Consequently, ASD causes lifelong disabilities on the individuals who are suffering from and inserts significant burdens on their families, schools, and society (Wazana *et al.*, 2007). It is however believed that environmental factors and heritability contribute to autism, researchers anticipating that genetic factors are playing a major role in the occurrence of the disorder (Thurm and Swedo, 2012). Nevertheless, it was suspecting that the environmental factors, such as vaccine hypothesis, can be another cause of autism, but this was not truly approved and hence

the author of that study has taken away his license of physician (Alarcón *et al.*, 2008). Supporting to the genetic causes, recent studies showed the presence of a high similarity in the genetics of autistic among the genetic features of autistic twin (Taniai *et al.*, 2008). Interestingly, it was observed that the genetic similarity is high among identical twins who are from the same developmental environment and same parental chromosomes.

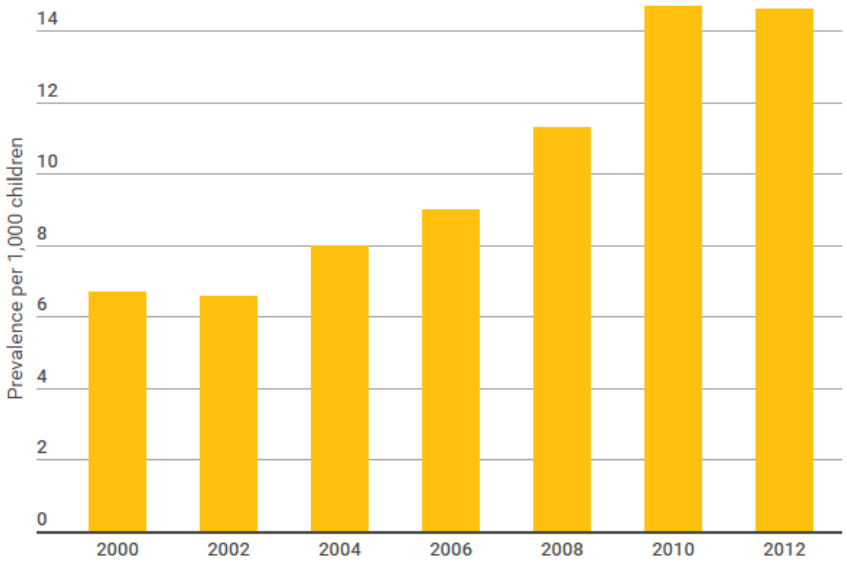

Figure 1.1: Autism prevalence during 2000 to 2012 (MOISSE, 2016).

The genomics science is the field of studying genes and their functions. Genomics provide a great approach to perform interesting researches on autism as it facilitates investigation on the global changes of gene expression (Gregg *et al.*, 2008). Each gene is analogous to a chapter of an instruction book, revealing the theory behind creating a specific family of molecules. As such, the genes that are known as protein-coding identifies how to establish large molecules that are made from amino-acid chains (proteins) while that the non-coding genes define the way of creating small molecules that are made from ribonucleic acid (RNA) chains (Leung *et al.*, 2016). Therefore, investigation on the expression level of genes for both healthy and non-healthy samples helps in recognizing the altered gene expression of particular genes. In these contexts, biologists require to identify the most relevant genes that can be utilized as biomarkers for tracing a known disease.

Consequently, the attributed genes enable us to understand the formation mechanism of the disease as well as predicting the serious danger of such disease. Up until now, there is a lack of treatment for the major symptoms of ASD and even no accurate biomarkers have been identified to diagnose this disorder (Yoo, 2015). This is basically due to the fact that the etiology of autism is not clearly known yet. Nevertheless, the possibility of heritability causes is expected to be about 70% to 90%, yet the changeable phenotype and complex architecture of genetic has become a bottleneck in front of identifying the susceptible genes of autism (Alarcón *et al.*, 2008). It has been claimed that the aggregative actions of multiple genes are responsible to produce autism disorder, which in turn gives more complexity to the disorder in terms of genomic investigations (Purcell *et al.*, 2001). The pioneer work of Gregg *et al.* (2008), which was made upon using genomic profiling of whole blood has shown differences between gene expression among the autistic and healthy children. Besides, they observed variations in gene expression between the subtypes of autism such as autism with regression and without regression at the early onset stage. Because of these variations, the identification of most related genes to autism disorder has become a common problem. This is not a challenging task for biologists only, but for computer scientists as well, especially when building a generalized model for autistic genes is targeted.

Moreover, it is quite reasonable to use gene expression data to relate the phenotypes of disease and its attributed biomarkers (de Menezes *et al.*, 2004; Leung and Cavalieri, 2003). Interestingly, computer models can be effectively applied to recognize autism through using the microarray data of gene expression (Hu and Lai, 2013; Stahl *et al.*, 2012). Microarray is a tool used to estimate whether mutation in genes has occurred for a particular individual. The information is recorded in a microarray chip, which consists of a small glass plate enclosed in plastic. Fabrication of microarrays by some companies is almost similar in methodology to those of the production of computer microchips. The surface of each chip comprises thousands of short, synthetic and single-stranded DNA sequences (Govindarajan *et al.*, 2012). These are added together to the investigated normal gene and to its variants (mutations) which have been observed in the human population. The techniques of machine

learning and datamining are considered as an effective tool in the application of genomic medicine that depend on computational problems and datasets in order to predict phenotypes (Leung *et al.*, 2016). These techniques are dealing with the development and employment of statistical methods as well as machine learning algorithms which are capable to be improved with experience. Machine learning is mostly important to interpret large datasets of genomic, it can also be successfully utilized to annotate a wide diversity of elements in genomic sequence (Libbrecht and Noble, 2015).

Studies in cancer informatics has shown great contribution of datamining and machine learning in finding the related genes (Chandra Sekhara Rao Annavarapu and Banka, 2016; Guyon *et al.*, 2002; Rejani and Selvi, 2009) since abnormalities in genes is led to alteration in the gene expression values of those genes. These studies have generically found the very specific genes related with each type of cancer and based on this, various models have been proposed in the literature to predict the risk of having cancer (Alba *et al.*, 2007; Tran *et al.*, 2014). However, autism gene expression data is having some specific characteristics, which make the feature selection, model creation and prediction more challenging than cancer gene expression analysis. As mentioned earlier, the contribution of many genes and environmental factors in the appearance of this spectrum disorder has made the gene expression profile of autistic people to be characterized by high variance in many genes rather than a group of genes. However, these variances could not be seen equally in different types and different groups of autisms. They may appear in one gene for an individual, while they may appear in many genes for another one (Yoo, 2015).

It is worth to mention that studies about autism gene expression are few, most of the studies have been done in behaviour detection and autism neurology (Beacher *et al.*, 2012; Crippa *et al.*, 2015; Elsabbagh *et al.*, 2009). Due to the importance of finding those genes related to autism (Muhle *et al.*, 2004; Taniai *et al.*, 2008; Yoo, 2015) computer models can take part in this finding. This can be achieved by building some models consisting of statistic and machine learning algorithms aiming at handling the high variance and dealing with many altered genes. Figure 1.2 shows the

representation of existed relation between autism and gene alteration, which may include, deletion, duplication and inversion.

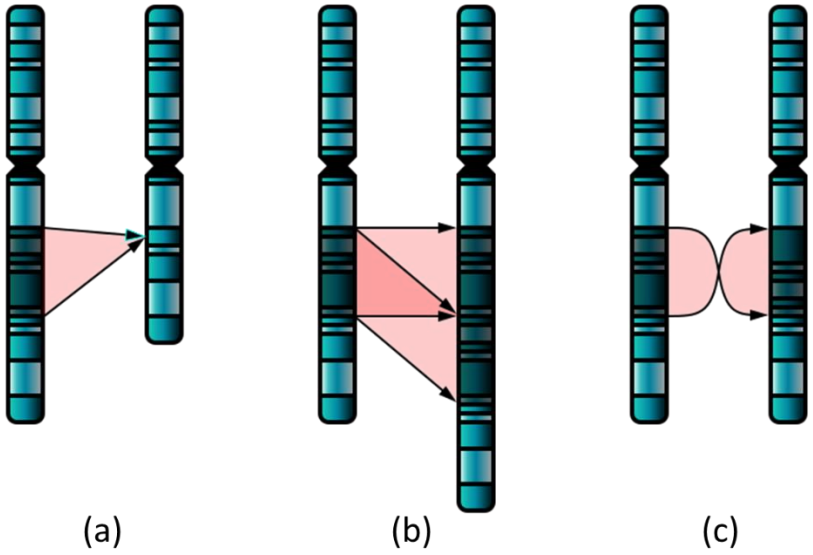

Figure 1.2: Deletion (a), duplication (b) and inversion (c) are all chromosome abnormalities that have been implicated in autism (Beaudet, 2007).

1.2 Challenges in Autism Microarray Data Analysis

The data of autism gene expression encounters the problem of limited number of observations (samples) compared to the high number of features (genes). For instance, the dataset which is used in the current study contains very high features of more than 54,600 genes, whereas the number of observations are limited to only 146 samples. This in turn results in a high imbalance between the number of genes and observations (patients). Dealing with such data needs more attention and it requires a sophisticated model, such that it can handle such high number of features or genes.

Another existed challenge is due to the lack of enough data to be used in studies on autism disorder. There are few datasets related to ASD which are available through the well-known NCBI repositories (database, 2011), while each of them represents different data for a specific type of autism, and some of them represent different

component of genes such as RNA-seq and protein structure (Crippa *et al.*, 2015; Nishimura *et al.*, 2007; Release, 2016). Moreover, biologically, the group of autism disorder presents different gene alterations in both number and in type (Lenroot and Yeung, 2013).

Therefore, it would be difficult to get the exact biomarker genes that are responsible for the disorder unless very careful investigations are made or there should be more datasets to be available for that purpose. As autism is recognized by a broad spectrum of disorder (Thurm and Swedo, 2012) its biology is different enough from those of the cancer and other diseases. Having various causes to produce the disorder and many different genes to contribute requires serious efforts to be made in order to find those genes which are related to a specific type of autism. This could be mainly due to the lack of previous works on analysing a specific type of autism and the nonexistence of datasets for similar group of people having the same phenotypes. As explained previously, due to the wide range of gene contribution to autism, their gene expression values in the current datasets have the problem of high variance among so many genes which can be found clearly in autism group.

For example, in the dataset of autism downloaded from NCBI database (database, 2011), the variance of gene expression values of different individuals changing from 0.099 to 24.19 $\times 10^6$ with the mean of 2.38 $\times 10^4$, median of 38.84 and 13,245 genes of the expression variance higher than 1000. Nevertheless, a portion of variance can also be seen in control (normal) group, but this kind of variance is related to the noise, which may be resulted from having more than one batch or the devices which are used to generate gene expression. Therefore, giving extra efforts to build a computerized model that can identify the biomarker genes for autism spectrum disorder and classify autistic from non-autistic samples are of primary request in this study. This will be done via applying some steps of statistical operations and three filters in parallel, followed by a wrapper feature selection then combining the results of feature selection branches to one ensemble form.

1.3 Problem Statement

The main problem in gene expression analysis is the difficulty of selection and identification of most relevant/biomarker genes to autism. This is due to the presence of limited number of observations in comparison to the very large number of genes, which is known as high dimensionality, requiring sophisticated methods to handle it.

Moreover, the datasets which are generated by microarray technology have large number of gene expression values, leading to the complexity in terms of datamining and machine learning analysis. Furthermore, they contain noise, which is defined as "the error in the variance of a measured variable", resulting from errors in measurements or natural variation (Han *et al.*, 2011; Hira and Gillies, 2015). Determination of the real expression values from the noise is one of the main problems in gene expression analysis.

Furthermore, the gene expression levels in autism disorder are highly noisy and several sequences of these genes show a large variance. In autism, the extra variance may be linked with alteration in many genes. Consequently, it is not an easy task to find the attributed genes straightforward unless careful analysis and investigation are made upon the microarray dataset through the application of various criteria and algorithms during the pre-selection, selection and classification of the expressed genes.

Hence, this thesis is expected to answer the following questions:

(i) How to reduce the high dimensionality of gene expression dataset by removing the most similar genes and pre-selection?

(ii) How to effectively find the most important genes despite the presence of high variance and noise in the dataset?

(iii) How can a model be generated based on the selected genes to make prediction on the autism dataset?

(iv) Does an ensemble of some input methods can make the performance of the model more accurate?

1.4 Objective of the Study

The objective of the current research work can be summarized as follows:

(i) To propose statistical based pre-selection and three filter methods in dealing with high variance and redundancy in autism gene expression dataset and for gene selection accordingly.

(ii) To apply a hybrid form of machine learning algorithm, known as geometric binary particle swarm optimization-support vector machine (GBPSO-SVM) in the second phase of selection process aiming at effective achievement of accurate features representing discriminant genes for the autism disorder.

(iii) To address the accuracy of the utilized models in each step and using ensemble classification (Random Forest (RF) and/or SVM) by combining the results of different outputs from final selection steps, thereby producing a fruitful result.

1.5 Scopes and Limitations

The contribution of the current work is specifically limited to the field of gene expression analysis of Autism Spectrum Disorder (ASD). This is aiming at identifying the most relevant genes to autism such that they can be used as biomarkers for the diagnosis of the disorder. Besides, the built model could be used to predict and recognize the risk of existing autism or not. Hence, the scopes of this research are illustrated as follows:

(i) The experimental procedures and analysis are performed on Windows 64-bit Operating system, Corei5, 2.7 GHz processor speed and 8 GB RAM.

(ii) MATLAB programming (R2016a) and Weka programming (version 3.8) are used.

(iii) The database benchmark in this work is related to autism disorder.

(iv) The dataset is consisting of 54613 features/genes and 146 observations/samples.

(v) The dataset format is in (. soft) tab delimited text file, which is a special text format used by NCBI-GEO database repository (database, 2011).

(vi) Multiple criteria based on a combination of mean, median and variance are used followed by the utilization of three filter methods, namely two sample t-test (T-test), Wilcoxon rank sum (WRS), and feature correlation with class (COR) test, in the process of gene selection.

(vii) The applied wrapper form of machine learning method is in the form of hybrid (GBPSO-SVM), which is used as the second phase of feature selection. RF algorithm is also employed in parallel with SVM at the final step of the classifications to represent the ensemble of the fused results.

1.6 Significance of the Research

The significance of this study is two-fold, which can be categorized into computational and biological aspects. From the computational approach, the proposed method is aimed at finding the most related genes to autism then making classification upon them. Moreover, increasing the efficiency of the model by proposing new statistical and machine learning criteria of gene selection. What is related to the biological approach is that the results of this study could help the medical and

biological sectors to further investigate those genes which have been identified during this research.

1.7 Thesis Outline

In Chapter one, an introduction and statement of the thesis is given which includes a general background on the topic and problem statement followed by challenges in the data analysis of autism gene expression, and objective of the thesis. As well as, the scope and significance of the study are presented therein along with the thesis outline as its last section. Chapter two reports a literature review on the methods and techniques used for the reduction of dataset, gene feature selection and classification processes, while Chapter three illustrates the proposed methodology of the research work in detail. In Chapter four, the obtained results are analysed and discussed. Finally, Chapter five is devoted to draw the conclusions and future suggested works.

REFERENCES

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics, 26*(3), 392-398.

Alarcón, M., Abrahams, B. S., Stone, J. L., Duvall, J. A., Perederiy, J. V., Bomar, J. M., Sebat, J., Wigler, M., Martin, C. L., and Ledbetter, D. H. (2008). Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *The American Journal of Human Genetics, 82*(1), 150-159.

Alba, E., Garcia-Nieto, J., Jourdan, L., and Talbi, E.-G. (2007). *Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms.* Paper presented at the Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 284-290.

Alter, M. D. (2013). Autism and Increased Paternal Age Related Changes in Global Levels of Gene Expression Regulation. Public Library of Science ONE Journal: Februari.

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., and Jones, J. J. (2011). Autism and increased paternal age related changes in global levels of gene expression regulation. *PloS one, 6*(2), e16715.

Ardjani, F., Sadouni, K., and Benyettou, M. (2010, 27-28 Nov. 2010). *Optimization of SVM MultiClass by Particle Swarm (PSO-SVM).* Paper presented at the 2010 2nd International Workshop on Database Technology and Applications, 1-4.

Beacher, F. D., Radulescu, E., Minati, L., Baron-Cohen, S., Lombardo, M. V., Lai, M.-C., Walker, A., Howard, D., Gray, M. A., and Harrison, N. A. (2012). Sex differences and autism: brain function during verbal fluency and mental rotation. *PLoS One, 7*(6), e38355.

Beaudet, A. L. (2007). Autism: highly heritable but not inherited. *Nature medicine, 13*(5), 534-537.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of computational biology, 7*(3-4), 559-583.

Ben Ayed, A., Benhammouda, M., Ben Halima, M., and Alimi, A. M. (2017). *Random forest ensemble classification based fuzzy logic*, 103412B-103412B-103415.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems, 34*(3), 483-519.

Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences, 282*, 111-135.

Bonyadi, M. R., and Michalewicz, Z. (2017). Particle swarm optimization for single objective continuous space problems: a review: MIT Press.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Bridge, P. D., and Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of clinical epidemiology, 52*(3), 229-235.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, 22.21. 21-22.21. 11.

Butler-Yeoman, T., Xue, B., and Zhang, M. (2015). *Particle swarm optimisation for feature selection: A hybrid filter-wrapper approach.* Paper presented at the Evolutionary Computation (CEC), 2015 IEEE Congress on, 2428-2435.

Carlos J. Alonso-González, Q. I. M.-S., Arancha Simon-Hurtado, Ricardo Varela-Arrabal. (2012). Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications*.

Cateni, S., Vannucci, M., Vannocci, M., and Colla, V. (2013). Variable Selection and Feature Extraction Through Artificial Intelligence Techniques. In L. V. d. Freitas and A. P. B. R. d. Freitas (Eds.), *Multivariate Analysis in Management, Engineering and the Sciences* (pp. Ch. 06). Rijeka: InTech.

Cervante, L., Xue, B., Zhang, M., and Shang, L. (2012). *Binary particle swarm optimisation for feature selection: A filter based approach.* Paper presented at the Evolutionary Computation (CEC), 2012 IEEE Congress on, 1-8.

Chandra Sekhara Rao Annavarapu, S. D., and Banka, H. (2016). Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. *EXCLI journal, 15*, 460.

Chen, L.-F., Su, C.-T., Chen, K.-H., and Wang, P.-C. (2012). Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. *Neural Computing and Applications, 21*(8), 2087-2096.

Chen, Y., Miao, D., and Wang, R. (2010). A rough set approach to feature selection based on ant colony optimization. *Pattern Recognition Letters, 31*(3), 226-233.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cosma, G., Brown, D., Archer, M., Khan, M., and Pockley, A. G. (2017). A survey on computational intelligence approaches for predictive modeling in prostate cancer. *Expert Systems with Applications, 70*, 1-19.

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., and Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of autism and developmental disorders, 45*(7), 2146-2156.

Das, S. (2001). *Filters, wrappers and a boosting-based hybrid for feature selection.* Paper presented at the ICML, 74-81.

database, N. (2011). Autistic children and their father's age: peripheral blood lymphocytes (Publication., from from www.ncbi.nlm.nih.gov: http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431

de Menezes, R. X., Boer, J. M., and van Houwelingen, H. C. (2004). Microarray Data Analysis. *Applied Bioinformatics, 3*(4), 229-235.

De Rubeis, S., and Buxbaum, J. D. (2015). Recent advances in the genetics of autism spectrum disorder. *Current neurology and neuroscience reports, 15*(6), 1-9.

Dehmer, M., and Emmert-Streib, F. (2008). *Analysis of microarray data: A network-based approach*: John Wiley & Sons.

Deng, L., Pei, J., Ma, J., and Lee, D. L. (2004). *A rank sum test method for informative gene discovery.* Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 410-419.

Derrick, B., Toher, D., and White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods in Psychology, 12*(1), 30-38.

Díaz-Uriarte, R., and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics, 7*(1), 3.

Duda, R. O., and Hart, P. E. (2003). *Pattern Classification and Scene Analysis* New York, NY: John Wiley.

Ebrahimpour, M. K., and Eftekhari, M. (2017). Ensemble of feature selection methods: A hesitant fuzzy sets approach. *Applied Soft Computing, 50*, 300-312.

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

Elsabbagh, M., Volein, A., Csibra, G., Holmboe, K., Garwood, H., Tucker, L., Krljes, S., Baron-Cohen, S., Bolton, P., and Charman, T. (2009). Neural correlates of eye gaze processing in the infant broader autism phenotype. *Biological psychiatry, 65*(1), 31-38.

Fisher, R. A. (1925). *Theory of statistical estimation.* Paper presented at the Mathematical Proceedings of the Cambridge Philosophical Society, 700-725.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *science*, 767-773.

Fong, S., Zhuang, Y., Tang, R., Yang, X.-S., and Deb, S. (2013). Selecting optimal feature set in high-dimensional data by swarm search. *Journal of Applied Mathematics, 2013*.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906-914.

García-Nieto, J., Alba, E., Jourdan, L., and Talbi, E. (2009). Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters, 109*(16), 887-896.

Geschwind, D. H., and State, M. W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. *The Lancet Neurology, 14*(11), 1109-1120.

Gnana, D. A. A., Appavu, S., and Leavline, E. J. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. *methods, 136*(1).

González, F., and Belanche, L. A. (2013). Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy. *arXiv preprint arXiv:1302.1733*.

Govindarajan, R., Duraiyan, J., Kaliyappan, K., and Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy & Bioallied Sciences, 4*(Suppl 2), S310-S312.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., and Initiative, A. s. D. N. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage, 65*, 167-175.

Gregg, J. P., Lit, L., Baron, C. A., Hertz-Picciotto, I., Walker, W., Davis, R. A., Croen, L. A., Ozonoff, S., Hansen, R., and Pessah, I. N. (2008). Gene expression changes in children with autism. *Genomics, 91*(1), 22-29.

Gunasundari, S., Janakiraman, S., and Meenambal, S. (2016). Velocity Bounded Boolean Particle Swarm Optimization for improved feature selection in liver and kidney disease diagnosis. *Expert Systems with Applications, 56*, 28-47.

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46*(1), 389-422.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Har-Peled, S., Roth, D., and Zimak, D. (2002). Constraint classification for multiclass classification and ranking. *Urbana, 51*, 61801.

Hassan, R., Cohanim, B., De Weck, O., and Venter, G. (2005). *A comparison of particle swarm optimization and the genetic algorithm.* Paper presented at the 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 1897.

Hassanien, A. E., Al-Shammari, E. T., and Ghali, N. I. (2013). Computational intelligence techniques in bioinformatics. *Computational biology and chemistry, 47*, 37-47.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one, 6*(12), e28210.

He, F., Yang, H., Wang, G., and Cui, G. (2012). *A novel method for hepatitis disease diagnosis based on RS and PSO.* Paper presented at the Proc. of International Conference of 4th Electronic System-Integration Technology Conference, 1289-1292.

Hira, Z. M., and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics, 2015*.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, 278-282.

Hoseini, P., and Shayesteh, M. G. (2013). Efficient contrast enhancement of images using hybrid ant colony optimisation, genetic algorithm, and simulated annealing. *Digital Signal Processing, 23*(3), 879-893.

Hu, V. W., Frank, B. C., Heine, S., Lee, N. H., and Quackenbush, J. (2006). Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC genomics, 7*(1), 1.

Hu, V. W., and Lai, Y. (2013). Developing a predictive gene classifier for autism spectrum disorders based upon differential gene expression profiles of phenotypic subgroups. *North American journal of medicine & science, 6*(3).

Huerta, E. B., Duval, B., and Hao, J.-K. (2006). *A hybrid GA/SVM approach for gene selection and classification of microarray data.* Paper presented at the Workshops on Applications of Evolutionary Computation, 34-44.

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

Hussain, M. A., Ansari, T. M., Gawas, P. S., and Chowdhury, N. N. (2015). Lung Cancer Detection Using Artificial Neural Network & Fuzzy Clustering. *International Journal of Advanced Research in Computer and Communication Engineering, 4*(3).

Jirapech-Umpai, T., and Aitken, S. (2005). Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC bioinformatics, 6*(1), 148.

Kar, S., Sharma, K. D., and Maitra, M. (2016, 28-30 Jan. 2016). *A particle swarm optimization based gene identification technique for classification of cancer subgroups.* Paper presented at the 2016 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC), 130-134.

Kennedy, J., and Eberhart, R. C. (1997). *A discrete binary version of the particle swarm algorithm.* Paper presented at the Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on, 4104-4108.

Khoshgoftaar, T., Dittman, D., Wald, R., and Fazelpour, A. (2012). *First order statistics based feature selection: A diverse and powerful family of feature seleciton techniques.* Paper presented at the Machine Learning and Applications (ICMLA), 2012 11th International Conference on, 151-157.

Kohavi, R., and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*(1-2), 273-324.

Koschmieder, A., Zimmermann, K., Trißl, S., Stoltmann, T., and Leser, U. (2012). Tools for managing and analyzing microarray data. *Briefings in bioinformatics, 13*(1), 46-60.

Krey, J. F., and Dolmetsch, R. E. (2007). Molecular mechanisms of autism: a possible role for Ca 2+ signaling. *Current opinion in neurobiology, 17*(1), 112-119.

Lai, C., Reinders, M., and Wessels, L. (2005). *Multivariate gene selection: does it help?* Paper presented at the Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE, 99-100.

Lai, C., Reinders, M. J., van't Veer, L. J., and Wessels, L. F. (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC bioinformatics, 7*(1), 235.

Latkowski, T., and Osowski, S. Gene selection in autism – Comparative study. *Neurocomputing*.

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

Latkowski, T., and Osowski, S. (2015b). Data mining for feature selection in gene expression autism data. *Expert Systems with Applications, 42*(2), 864-872.

Latkowski, T., and Osowski, S. (2015c). *Developing Gene Classifier System for Autism Recognition.* Paper presented at the International Work-Conference on Artificial Neural Networks, 3-14.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., and Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(4), 1106-1119.

Lenroot, R. K., and Yeung, P. K. (2013). Heterogeneity within Autism Spectrum Disorders: What have We Learned from Neuroimaging Studies? *Frontiers in Human Neuroscience, 7*, 733.

Leung, M. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE, 104*(1), 176-197.

Leung, Y. F., and Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *TRENDS in Genetics, 19*(11), 649-659.

Li, S., Wu, X., and Tan, M. (2008). Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing-A Fusion of Foundations, Methodologies and Applications, 12*(11), 1039-1048.

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*(6), 321-332.

Libralon, G. L., Carvalho, A. C. P. d. L., and Lorena, A. C. (2009). Pre-processing for noise detection in gene expression classification data. *Journal of the Brazilian Computer Society, 15*(1), 3-11.

Liu, M., Wang, M., Wang, J., and Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical, 177*, 970-980.

Liu, X., and Shang, L. (2013). *A fast wrapper feature subset selection method based on binary particle swarm optimization.* Paper presented at the Evolutionary Computation (CEC), 2013 IEEE Congress on, 3347-3353.

MOISSE, K. (2016). U.S. stats show autism rate reaching possible plateau. *SpectrumNews* Retrieved March 31, 2016, from

https://spectrumnews.org/news/u-s-stats-show-autism-rate-reaching-possible-plateau/

Moradi, P., and Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing, 43*, 117-130.

Moraglio, A., Di Chio, C., Togelius, J., and Poli, R. (2008). Geometric particle swarm optimization. *Journal of Artificial Evolution and Applications, 2008*.

Moraglio, A., Togelius, J., and Silva, S. (2013). Geometric differential evolution for combinatorial and programs spaces. *Evol. Comput., 21*(4), 591-624.

Muhle, R., Trentacoste, S. V., and Rapin, I. (2004). The genetics of autism. *Pediatrics, 113*(5), e472-e486.

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

Nishimura, Y., Martin, C. L., Vazquez-Lopez, A., Spence, S. J., Alvarez-Retuerto, A. I., Sigman, M., Steindler, C., Pellegrini, S., Schanen, N. C., and Warren, S. T. (2007). Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Human molecular genetics, 16*(14), 1682-1698.

Ozcift, A., and Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer methods and programs in biomedicine, 104*(3), 443-451.

Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences, 91*(11), 5022-5026.

Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics, 9*(1), S13.

Puissant, A., Rougier, S., and Stumpf, A. (2014). Object-oriented mapping of urban trees using Random Forest classifiers. *International Journal of Applied Earth Observation and Geoinformation, 26*, 235-245.

Purcell, A., Jeon, O., Zimmerman, A., Blue, M., and Pevsner, J. (2001). Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology, 57*(9), 1618-1628.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med, 5*(9), e184.

Rays, M., Chen, Y., and Su, Y. A. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature genetics, 14*.

Rejani, Y., and Selvi, S. T. (2009). Early detection of breast cancer using SVM classifier technique. *arXiv preprint arXiv:0912.2314*.

Release, O. N. (2016). Researchers flag hundreds of new genes that could contribute to autism (Publication., from Princton University: https://www.princeton.edu/main/news/archive/S47/03/24S01/index.xml?section=newsreleases

Russo, G., Zegar, C., and Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene, 22*(42), 6497-6507.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688-690.

Sadakata, T., Sekine, Y., Oka, M., Itakura, M., Takahashi, M., and Furuichi, T. (2012). Calcium-dependent activator protein for secretion 2 interacts with the class II ARF small GTPases and regulates dense-core vesicle trafficking. *The FEBS journal, 279*(3), 384-394.

Sadakata, T., Shinoda, Y., Ishizaki, Y., and Furuichi, T. (2017). Analysis of gene expression in Ca2+-dependent activator protein for secretion 2 (Cadps2) knockout cerebellum using GeneChip and KEGG pathways. *Neuroscience letters, 639*, 88-93.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. *Machine learning and knowledge discovery in databases*, 313-325.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.

Saha, S., Seal, D. B., Ghosh, A., and Dey, K. N. (2016). A novel gene ranking method using Wilcoxon rank sum test and genetic algorithm. *International Journal of Bioinformatics Research and Applications, 12*(3), 263-279.

Salem, H., Attiya, G., and El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing, 50*, 124-134.

Sandin, I., Andrade, G., Viegas, F., Madeira, D., Rocha, L., Salles, T., and Gonçalves, M. (2012). *Aggressive and effective feature selection using genetic programming.* Paper presented at the 2012 IEEE Congress on Evolutionary Computation, 1-8.

Singh, R. K., and Sivabalakrishnan, M. (2015). Feature selection of gene expression data for cancer classification: a review. *Procedia Computer Science, 50*, 52-57.

Spies, N. (2014). Machine Learning For Cancer Classification - Part 2 - Building A Random Forest Classifier. *BioStars*, https://www.biostars.org/p/86981/.

Sprent, P., and Smeeton, N. C. (2016). *Applied nonparametric statistical methods*: CRC Press.

Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., and Team, B. (2012). Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism. *Developmental neuropsychology, 37*(3), 274-298.

Stamou, M., Streifel, K. M., Goines, P. E., and Lein, P. J. (2013). Neuronal connectivity as a convergent target of gene× environment interactions that confer risk for Autism Spectrum Disorders. *Neurotoxicology and teratology, 36*, 3-16.

Talbi, E.-G., Jourdan, L., Garcia-Nieto, J., and Alba, E. (2008). *Comparison of population based metaheuristics for feature selection: Application to microarray data classification.* Paper presented at the Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, 45-52.

Taniai, H., Nishiyama, T., Miyachi, T., Imaeda, M., and Sumi, S. (2008). Genetic influences on the broad spectrum of autism: Study of proband-ascertained twins. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 147*(6), 844-849.

Thurm, A., and Swedo, S. E. (2012). The importance of autism research. *Dialogues on Clinical Neurosciences, 14*(3), 219-222.

Tran, B., Xue, B., and Zhang, M. (2014). *Improved PSO for feature selection on high-dimensional datasets.* Paper presented at the Asia-Pacific Conference on Simulated Evolution and Learning, 503-515.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-525.

Walt, D. R. (2000). Bead-based fiber-optic arrays. *Science, 287*(5452), 451-452.

Wang, X., Yang, J., Teng, X., Xia, W., and Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern recognition letters, 28*(4), 459-471.

Wang, Y., Wang, P., Xu, X., Goldstein, J., McConkie, A., Cheung, S. W., and Jiang, Y.-H. (2015). Genetics of Autism Spectrum Disorders: The Opportunity and Challenge in the Genetics Clinic. In *The Molecular Basis of Autism* (pp. 33-66): Springer.

Wazana, A., Bresnahan, M., and Kline, J. (2007). The autism epidemic: fact or artifact? *Journal of the American Academy of Child & Adolescent Psychiatry, 46*(6), 721-730.

Wikipedia. (2017). Messenger RNA. *Wikipedia the free encyclopedia,* https://en.wikipedia.org/wiki/Messenger_RNA.

Wild, C., and Seber, G. (2011). The Wilcoxon rank-sum test: Chapter.

Wiliński, A., and Osowski, S. (2012). Ensemble of data mining methods for gene ranking. *Bulletin of the Polish Academy of Sciences: Technical Sciences, 60*(3), 461-470.

Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft computing and industry* (pp. 25-42): Springer.

Wolpert, D. H., and Macready, W. G. (1995). *No free lunch theorems for search*: Technical Report SFI-TR-95-02-010, Santa Fe Instituteo. Document Number)

Wonnacott, T. H., and Wonnacott, R. J. (1972). *Introductory statistics* (Vol. 19690): Wiley New York.

Xu, M., and Setiono, R. (2015). Gene selection for cancer classification using a hybrid of univariate and multivariate feature selection methods. *arXiv preprint arXiv:1506.02085*.

Xue, B., Zhang, M., and Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE transactions on cybernetics, 43*(6), 1656-1671.

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

Zhao, S., Dong, X., Shen, W., Ye, Z., and Xiang, R. (2016). Machine learning-based classification of diffuse large B-cell lymphoma patients by eight gene expression profiles. *Cancer medicine*.

Zhou, L.-T., Cao, Y.-H., Lv, L.-L., Ma, K.-L., Chen, P.-S., Ni, H.-F., Lei, X.-D., and Liu, B.-C. (2017). Feature selection and classification of urinary mRNA microarray data by iterative random forest to diagnose renal fibrosis: a two-stage study. *Scientific Reports, 7*.

Zimmerman, D. W., and Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education, 62*(1), 75-86.