Making Medical Predictions about Diseases with Gene Expression Data

Xuanwei Li [1,*]

[1] College of Engineering and Applied Science, Stony Brook University, Stony Brook, NY, 11794, USA

* Corresponding author: xuanwei.li@stonybrook.edu

Abstract. Exploration and studies of human genes play a critical role for improving the healthcare and society development. By using previous data as input, machine learning enables software applications to forecast values more precisely and is a good tool to complete the task of predicting diseases with gene expression data. This paper provides a comprehensive review of studies regarding the combination of machine learning and gene expression analysis related to diseases. The main three applications are (a) the disease prediction: cancer detection and other diseases detection, (b) the control of cancers: the metastasis of cancer and the complete remission of cancers, and (c) the drug response prediction. The reviewed molding method in this paper mainly focus on Regressions, K nearest neighbor (KNN) and Support vector machine (SVM). The combination of gene data and machine learning is meaningful for developing new techniques for detecting diseases and testing new drugs, which improves accuracy and effectiveness.

Keywords: Machine learning, gene expression, detect diseases, drug response.

1. Introduction

Exploration and studies of human genes are closely related and widely applied to plenty of disciplines like healthcare, society resources, technology, and computational biology. One of the most significant applications is in the biomedical industry. Research of genes contributes to the development and invention of various medicine by combining human genes and disease features. Using different biomarkers and gene expression products, the industry of producing more efficient and effective medicine is developing rapidly. Furthermore, gene therapy is increasingly mature with deeper studies, which is a technique that uses a gene to treat, prevent or cure diseases or medical disorders. Another area that is worth mentioning is to detect and predict diseases with genes and their expression product. Not only most hereditary diseases are connected tightly to human genes, but many acquired diseases can be diagnosed with gene expression data.

Machine learning is a good tool to complete the task of predicting diseases with gene expression data. By using previous data as input, it is enables software programs to forecast values more precisely. [1]. Available large datasets of diseases provide a solid base for training the models with different methods and ensuring the basic accuracy of the final models. Researchers can train models to distinguish healthy gene expression products and ill ones of different diseases or classify data in another direction. For example, regressions like linear regression and logistic regression can be applied. There are also software packages to access the model and test the accuracy. The accuracy of models should reach a certain level to prove their effectiveness. Otherwise, researchers can fit data with other models to find the most suitable one.

This paper provides a comprehensive review of studies regarding the combination of machine learning and gene expression analysis related to diseases. The following is the main structure of the paper: some of the most popular models for machine learning will be introduced in section two, along with a great variety of methods and software packages. As for section three, the applications and research results in gene expression analysis concerning diseases with ML are provided by category. The main three applications are disease prediction, the prediction of control of cancers, and drug response prediction. As for disease prediction, there are two subsections, prediction for cancer and other diseases. For the control of cancer, two subsections of it are metastasis and complete remission

of cancer. Finally, in section four, the discussion of future development and conclusion are presented. The combination of gene data and machine learning is meaningful for developing new techniques for detecting diseases and testing new drugs, which improves accuracy and effectiveness.

2. Methods and models

2.1. Regressions

Regression is a supervised learning model that made use of existing data to foretell the value of fresh input. The most common regressions are linear regression and logistics regression. The main basic structure of regression is shown in Figure 1(a). As for linear regression, it builds a linear model for continuous variables. It finds the line which minimizes the cost function of data. The most common cost function for linear regression is the total squared residuals over all points, which is also called least squared regression. On the contrary, logistics regression is used for categorical variables. The function behind logistic regression is representing the likelihood that a particular sample belongs to a particular class. It also can get the final logistic regression model by optimizing the cost function, but the cost function is quite different from that of linear regression.

2.2. K nearest neighbor

K nearest neighbor (KNN) classifies categorical variables depends on the nearby items. K is a parameter representing how many nearest neighbors will be checked for the item like as shown in Figure 1(b). The vital part of KNN is the algorithms for calculating distance, which finds proper neighbors for items. There are also different algorithms for the process of voting for neighbors. For example, the item can simply be labeled with the label of most neighbors.

2.3. Clustering

Clustering [2] classifies items by similarity, which is a good way to reveal the origins of data from a small number of sources or explanations. The similarity of data is usually defined by some underlying distance function or metric. Circular clusters are natural but not universal. The k-means technique [3], the hierarchical technique [4], and the expectation-maximization technique [5] are well-known clustering techniques. Figure 1(c) is an illustration of structure of k-means technique.

2.4. SVM classifiers

Support vector machine (SVM) [2] is a binary classifier used for regression, clustering, and classification. SVM represents items as points in space of a certain dimension and maps individual points with as wide as possible intervals. Then when the new point is mapped into the same space, its label or category is based on what side it lies in. Besides linear classifiers, SVM can also use the “kernel trick” to implement non-linear classification [6]. The basic structure diagram is shown in Figure 1(d).

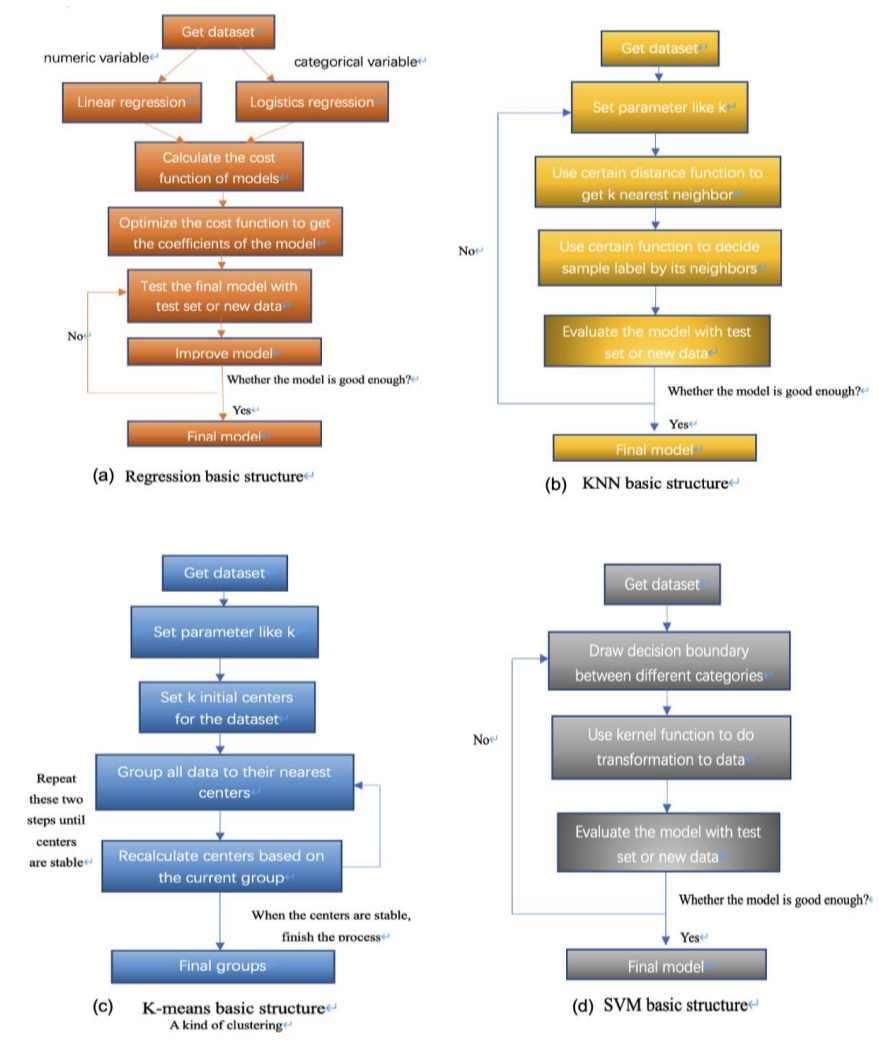


Figure 1. Basic structure of four machine learning models. (a) Regression, (b) KNN, (c) K-means, a kind of clustering, (d) SVM.

3. Reviews of studies

3.1. Diseases prediction

A timely diagnosis of the disease is an important step toward recovery. It is beneficial for preventing the progression of diseases, facilitating more effective treatment, and greatly reducing the probability of death. As the medical industry develops, techniques and technology for disease diagnosis have become increasingly mature and effective. This section is a review of various machine learning models that predict diseases from data on genes and their expression. Table 1 is an overview of research and their result regarding the prediction of diseases.

3.1.1 Predicting cancers

Genes and their expression product also contribute to the detection of tumors. For example, there are plenty of studies related to the important role of genes in predicting prostate cancer. Using algorithms like support vector machine (SVM) classifiers, Alkhateeb et al. [7] created a model for prostate cancer. They not only did experiments between healthy samples and ill ones, but also compared tissues among different stages of prostate cancer, as shown in Figure 2. As a result, they identified a series of genes and RNA sequences that show a strong correlation to prostate cancer, which can be used as biomarkers for predicting prostate cancer. Besides detecting the existence of tumors, these gene biomarkers are useful for determining the stage of cancer as well, which are significantly meaningful for practical clinical diagnosis. Furthermore, gene expression products can

also help to confirm the location of prostate cancer. In the research of Hamzeh, Osama, et al. [8], the prostate is divided into three parts, which are left unary, right unary, and bilateral. They worked on finding genes as biomarkers for each part with high accuracy. Using found genes, researchers can locate tumors and efficiently guide the following treatment. That kind of study may accelerate the replacement of the current inaccurate examination of cancer with a high missing up rate, leading to a revolution in the medical area.

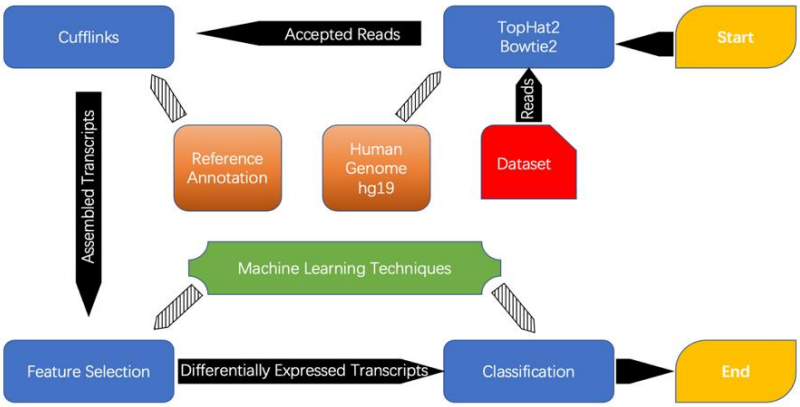


Figure 2. An illustration of the suggested methodology for identifying distinct transcripts in different prostate cancer stages and between benign and malignant tumors. [7].

Using genes to detect breast cancer is a hot topic as well. Breast cancer is one of the most prevalent and fatal diseases affecting women. Invasive ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS) are two kinds of it. Most studies focus on IDC because DCIS barely has negative effects on patients' daily lives and most breast cancer patients belong to the first type. What's more, it is not the first time for researchers to combine IDC prediction and machine learning. For years ago, researchers managed to use different models like CNN to classify and detect IDC with given patients' tissue images dataset. Researchers also made connections between genes and IDC. Roy et al. [9] created classification models regarding gene expression and IDC diagnosis. Applying patients' RNA sequences, the two stages of IDC, early and late, are separated using two-class classification models. They developed a web server called Duct-BRCA-CSP based on result models that have good accuracy in identifying the stage of IDC given patient samples. That study is also a robust support to the potential of genes in the diagnosis and treatment of cancers.

3.1.2 Other diseases

Besides cancer, there are other types of diseases that can be detected with genes. One of the non-cancer diseases that can be predicted with genes' help is osteoporosis (OP). Osteoporosis is a prevalent disease among the aging group all over the world. The early diagnosis of OP is beneficial to halt the progression, but the symptom of early osteoporosis is not obvious. Therefore, developing techniques for predicting or detecting OP become increasingly important [10,11]. Genes and their expression product can help to finish that job by working as biomarkers. Machine learning models like regressions and clusters were used by a team of researchers led by Chen, X., Liu, G. et al. [12] to correlate the gene characteristics of OP patients. By comparing the samples of OP and non-OP patients, they find out which gene or RNA plays a crucial part in the development of OP and then identify promising biomarkers for predicting OP. For a more detailed process, first, they did stepwise regression to select 22 genes that have more significant effects out of 50 key genes. Then, with further study and modeling, 9 genes are picked to build the final model. They are still significant in the analysis of multivariable logistic regression. Besides, all of them have ideal linear correlations with others. Hence, a relatively strong model can be built with 9 genes to detect OP in advance.

Another non-cancer example is the autism spectrum disorder (ASD). Depending on ASD's character, earlier diagnosis and treatment are more likely to get a better outcome [13]. Therefore, it is essential to develop a technique that can detect ASD as early as possible. Besides, based on previous

studies, there is proof indicating ASD may be hereditary and gene-related [14,15,16], so it can be a good choice to establish a gene-related predicting model for ASD. By comparing ASD children and the controlled group, some studies have made predicting models with blood-derived gene expression signatures for young children under 10. Dong Hoon Oh, et al [17] focused on the group of young adult patients. They aimed for finding biomarkers of ASD with blood samples. They built classifiers and predictive models, and then evaluated the performance of models. When using SVM and KNN analysis, the evaluation result is acceptable with an accuracy of around 98.8%. Nevertheless, the LDA analysis offers only an accuracy of 0.688, which is hugely different from previous analyses. Therefore, more studies and better models for predicting ASD among young adults with genes are needed.

Table 1. Main articles of prediction of diseases

Article	Diseases	Methods and models	Results
Alkhateeb A, et al. [7]	Prostate cancer	SVM classifiers	Finding plenty of genes as biomarkers for prostate cancer
Hamzeh, Osama, et al. [8]	Prostate cancer	SVM classifiers, random forest	identify groups of differentially expressed genes that can distinguish between samples belonging to the three classifications.
Roy, Shikha, et al. [9]	IDC	Random forest, regression, NB, decision tree and SVM	Classification models to determine whether IDC is in an early or late stage; Duct-BRCA-CSP
Chen, Xinlei, et al [12]	Osteoporosis	Regressions, clusters, PPI network	predicting model with 9 key genes
Dong Hoon Oh, et al. [17]	ADC	Cluster, SVM, KNN, LDA analysis	A not perfect model for predicting ADS among young adult

3.2. The control of diseases

It is mentioned above that genes and their expression product can diagnose different diseases and tumors with the combination of machine learning algorithms. In this part, it will be discussed the applications of genes in detecting tumor metastasis and complete remission, which also are significant for clinical practice and offer more chances for various treatments. Table 2 is the review of main articles mentioned concerning the control of diseases.

3.2.1. The metastasis of cancer

First, researchers paid attention to the applications of genes and their expression in detecting tumor metastasis. One example is the study of Seokjin Haam et al. [18], who attempted to utilize artificial intelligence (AI) to predict brain metastases in lung adenocarcinoma (LUAD) using gene expression signature. As the most prevalent non-small-cell lung cancer (NSCLC), LUAD frequently spreads to brain. The survival rate of LUAD patients with brain metastasis is quite low, where the median overall survival time of treated patients is only 4 -15 weeks. Therefore, biomarkers that can predict brain metastasis were needy. Seokjin Haam et al. focused mainly on the tumor nonimmune microenvironment and evaluated relative gene expression signatures. With different classifiers, the team finally found 17 gene expression signatures that are highly correlated to brain metastasis and can be used as predictive biomarkers. The above finding made it possible to detect brain metastasis of LUAD patients timelier, and it is also possible to make a specific and suitable plan for individual patients, especially for patients who have brain metastasis. What's more, another example is the study of Ye QH et al. [19], who developed models regarding hepatocellular carcinoma (HCC), which is a prevalent and high-fatality-rate disease. It is worth noticing that intra-hepatic metastases are mostly responsible for HCC's mortality rate. Hence, the research team of Ye QH et al. did a study regarding biomarkers of metastatic HCC with samples of patients and their gene expression profiles. They discovered the expression signatures of primary HCC with metastasis is highly comparable to those of their metastasis, demonstrating some initial tumor genes may be involved in the formation of metastasis. Then they found a lead gene named osteopontin which is over-expressed in samples with metastasis, indicating a strong connection with metastatic HCC. Therefore, osteopontin can be

considered a diagnostic biomarker for detecting metastatic HCC in advance, of great medical significance.

3.2.2. The complete remission of cancer

When it comes to complete remission, it is found that genes and their expression can also play a role in that area. Ophir Gal et al. [20] predicted the likelihood that individuals with acute myeloid leukemia (AML) would experience a full remission using the data of RNA sequencing. They compared sample patients who achieved complete remission after two courses of induction therapy and samples who did not turn to complete remission. Using different methods, classifiers and optimizers, researchers built different models and find the one with the best performance, which turns out to be the k-nearest neighbors' algorithm (KNN) model. By doing further tests to pick 50 genes, they finally found the essential role of N-glycosylation.

Table 2. Main articles of the control of cancer

Articles	Disease	Method and model	Results
Haam, Seokjin, et al. [18]	Brain metastasis of LUAD	Random forest, SVM, Naive Bayes, NN	17 genes can be used as predictive biomarkers
Ye QH et al. [19]	Metastatic HCC	4 different models	Key biomarker, osteopontin
Gal, Ophir, et al. [20]	AML	KNN models	N-glycosylation as biomarkers

3.3. Drug response prediction

Another top topic of genes and their expression is related to drug response. Drug reactions have always been the focus of academia. It is crucial for both clinical and medical practice, helping doctors to decide which drug treatment to use for specific individuals. Drug response also makes a difference in the progress of modern pharmaceutics by affecting the research and development of new drugs. Traditional ways of testing new drug responses have plenty of drawbacks. It needs considerable new drug samples and many Vivo experiments to test the effectiveness and side effects of that drug, which is expensive and needs a lot of human and material resources. What's more, the result may still not be accurate and comprehensive, leading to unexpected negative reactions. The development of machine learning and enriched databases offers a platform to revolute the way of testing drug response. There have been various developing techniques combining biochemistry and machine learning to test new drugs. In this part, machine learning models that use data of genes and their expression to predict drug responses are introduced.

There are plenty of studies aimed to establish general machine learning models to predict drug responses, meaning they are not only predicting the response of one specific drug. For example, Zuo, Zhaorui, et al. [21] built a deep learning model named SWnet, which applied genomic signatures to predict drug responses. Researchers can enter information on genetic mutation and expression signature into a gene weight layer in SWnet. One important feature of SWnet is that it investigates the relationship between genetics and drug chemical structure with a variety of machine learning models like GNN and CNN, giving a better prediction of drugs' effectiveness. Figure 3 showed a general illustration of GNN structure of SWnet. Based on its outstanding performance, SWnet can be used for drug screenings and is a promising model for cancer treatment.

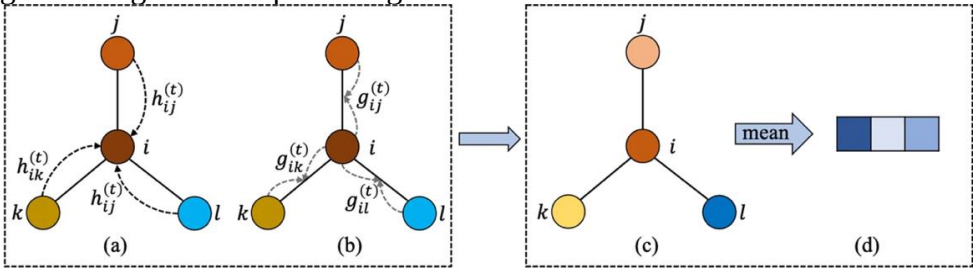


Figure 3. A general illustration of GNN structure. (a) Upgrade node embedding via nearby nodes. (b) Using side node embedding to upgrade embedding. (c) The graph following the transition function. (d) The arithmetic means of all nodes showing by graph [21]

Another model that worthies mentioning is the model proposed by Tan Mehmet [22]. He developed a model aimed to forecast how cancer cell lines will react to various drugs. Depending on recent pharmacogenomic databases, the algorithm uses multi-task learning and kernels to enhance its performance., and it turns out to be an excellent contender for screening anti-cancer drugs. In addition, from genomic information, Gönen M and Margolin AA [23] wrote algorithms for screening drugs for Human immunodeficiency virus (HIV) and cancer, which badly need individual treatments. Due to the fact that most existing algorithms did a model for each individual drug, they managed to take advantage of similarities between subsets of medications to create prediction models, using a unique Bayesian technique that combines binary classification with kernel-based non-linear dimensionality reduction. The study of a panel of drug-using multi-task learning finally gave a better model compared to models only concentrated on a single drug.

Researchers also built prediction models for specific drugs. For instance, He Manrong et al. [24] created a model for the response of ustekinumab (UST) for patients of Crohn's disease (CD) under treatment based on data from gene transcription profiling. UST is not the main treatment of CD, but when traditional therapies are not well responded to or they bring severe side effects to some patients, UST becomes a preferred alternative. With the Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene otology (GO), He Manrong et al. created a model that predicts UST response using regression models and got a successful prediction of test data.

4. Discussion

The application of data on genes and their expression opens a new way for the development of the medical industry. For predicting different diseases, the technique of using gene data gives a more accurate and direct prediction model. For some diseases, detecting them in advance can hugely increase the effectiveness of corresponding treatments, but it is hard to predict them. Gene provides a new angle for researchers to solve this problem, helping people diagnose some diseases early by combining them with machine learning algorithms. Besides, genes can also offer helps in detecting metastasis of different cancers and whether a patient is in complete remission of cancers. In this way, doctors can make a specific cancer treatment plan based on different individuals as early as possible, significantly improving the survival rate of patients. Turning to pharmaceutics, it is a revolution to predict drug response using studies of genes combined with machine learning. Traditional ways of testing drugs usually need a large amount of material and human resources, which is expensive and time-consuming. What's more, the completeness of datasets and the accuracy of test results cannot always be guaranteed. Using data of genes and their expression product accompanying machine learning can establish accurate and efficient models and systems to predict the response of various drugs. It not only accelerates the pace of new drug innovation, but also improves the techniques of drug screening, both of which make a difference in the development of pharmaceutics and medicine.

Nevertheless, there are limitations in this area. First, it is needy to gather considerable data to build a reliable and accurate model, but it is not an easy task to get a qualified dataset. Collecting and cleaning gene-relevant data from patients may be expensive and time-consuming, and the quality of data decides the quality and accuracy of the final models. Besides, it is impossible for rare diseases to get a large enough database to fit them into machine learning models. Therefore, researchers only are capable to use gene-related models in several diseases which have enough sample data. Second, not all disease indicates a relationship with their gene and gene expression. Although researchers with high-qualified data on a certain disease, the final model may show a weak correlation between gene and the disease, making it impossible to make a gene-related model.

To solve the limitation mentioned above, it is essential to establish and improve the medical information system to have more data. What's more, most diseases are not connected with the examination of genes and their expression, which also leads to the lack of data and should be paid attention to.

5. Conclusion

In conclusion, machine learning models can be combined with gene expression data, producing significant impacts on the development of the biomedicine industry. This article is a review of medical predictions of diseases with gene expression data, including the prediction of diseases, the control of cancer and drug response. This direction of study improves the accuracy and effectiveness of detecting different diseases and accelerates the development of medicine and pharmaceutics. The models using gene and machine learning provide a new angle for developing techniques for dealing with diseases and testing medicines. This review concludes some significant articles in this area to give a meaningful overview for future further studies.

References

[1] Burns, E., "What is machine learning and why is it important?", SearchEnterpriseAI, (2021).

[2] Liakos, K. G., Busato, P, Moshou, D., Pearson, S., and Bochtis, D., "Machine Learning in Agriculture: A Review," Sensors, (2018).

[3] Lloyd, S.P. "Least Squares Quantization in PCM," IEEE Trans. Inf. Theory, pg129-137 (1982).

[4] Johnson, S.C., "Hierarchical clustering schemes," Psychometrika, pg241-254(1967).

[5] Dempster, A.P., Laird, N.M. and Rubin, D.B, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B Methodol, pg1-37(1977).

[6] Vapnik, V., "Support vector machine," Mach. Learn, pg273-297, (1995).

[7] Alkhateeb, A., Rezaeian, I., Singireddy, S., Cavallo-Medved, D., Porter, L.A. and Rueda, L., "Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer," Cancer Informat. (2019).

[8] Hamzeh, O., Alkhateeb, A., Zheng, J., Kandalam, S., and Rueda, L., "Prediction of Tumor Location in Prostate Cancer Tissue Using a Machine Learning System on Gene Expression Data." BMC Bioinformatics, pp. 1–10 (2020).

[9] Roy, S., Kumar, R., Mittal, V., and Gupta, D., "Classification Models for Invasive Ductal Carcinoma Progression, Based on Gene Expression Data-Trained Supervised Machine Learning." Scientific Reports, pp. 1–15 (2020)

[10] Eastell, R. and Szulc, P., "Use of bone trnover markers in postmenopausal osteoporosis," Lancet Diabetes Endocrinol, (2017).

[11] Ensrud KE and Crandall CJ. "Osteoporosis," Ann Intern Med, (2017).

[12] Chen, X., Liu, G., Wang, S., Zhang, H., and Xue, P. "Machine Learning Analysis of Gene Expression Profile Reveals a Novel Diagnostic Signature for Osteoporosis," Journal of Orthopaedic Surgery & Research, pp. 1–8 (2021).

[13] Warren, Z., McPheeters, M. L., Sathe, N., Foss-Feig, J. H., Glasser, A., and Veenstra-Vanderweele, J. "A systematic review of early intensive intervention for autism spectrum disorders," Pediatrics, (2011).

[14] Constantino, J. N., Zhang, Y., Frazier, T., Abbacchi, A. M., and Law, P., "Sibling recurrence and the genetic epidemiology of autism," Am J Psychiatry, (2010).

[15] Ozonoff, S, Young, GS, Carter, A, Messinger, D, Yirmiya, N, and Zwaigenbaum, L., "Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study," Pediatrics, (2011).

[16] Rosenberg, R. E., Law, J. K., Yenokyan, G., McGready, J., Kaufmann, W. E., and Law, P. A., "Characteristics and concordance of autism spectrum disorders among 277 twin pairs," Arch Pediatr Adolesc Med, (2009).

[17] Oh, D. H., Kim, I. B., Kim, S. H., and Ahn, D. H. "Predicting Autism Spectrum Disorder Using Blood-Based Gene Expression Signatures and Machine Learning." Clinical Psychopharmacology & Neuroscience, pp. 47–52, (2017).

[18] Haam, S., Han, J. H., Lee, H. W., and Koh, Y. W. "Tumor Nonimmune-Microenvironment-Related Gene Expression Signature Predicts Brain Metastasis in Lung Adenocarcinoma Patients after Surgery: A Machine Learning Approach Using Gene Expression Profiling." Cancers, (2021).

[19] Ye, Q. H., Qin, L. X., Forgues, M., He, P., Kim, J. W., Peng, A. C., Simon, R., Li, Y., Robles, A. I., Chen, Y., Ma, Z. C., Wu, Z. Q., Ye, S. L., Liu, Y. K., and Tang, Z. Y., Wang, X. W., "Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning," Nat Med, (2003).

[20] Gal, O., Auslander, N., Fan, Y., and Meerzaman, D., "Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression," Cancer Informatics, (2019).

[21] Zuo, Z., Wang, P., Chen, X., Tian, L., Ge, H., and Qian. "SWnet: A Deep Learning Model for Drug Response Prediction from Cancer Genomic Signatures and Compound Chemical Structures." BMC Bioinformatics, pp. 1–16 (2021).

[22] Tan, M. "Prediction of anti-cancer drug response by kernelized multi-task learning," Artif Intell Med, (2016).

[23] Gönen, M. and Margolin, A. A., "Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning," Bioinformatics, (2014).

[24] He, M., Li, C., Tang, W., Kang, Y., Zuo, Y., and Wang, Y., "Machine Learning Gene Expression Predicting Model for Ustekinumab Response in Patients with Crohn's Disease," Immunity, Inflammation & Disease, (2021).