An Overview on Advanced Genetic Disease Diagnosis and Prediction Techniques Using Genome Data

M. Anitha[1], Mahendran Radha[2*]

[1]Department of Bioinformatics, School of Life Sciences, Research Scholar, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India.

[2*]Department of Bioinformatics, School of Life Sciences, Professor & Head, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India. E-mail: mahenradha@gmail.com, hodbioinfo@velsuniv.ac.in

Abstract

A genetic disorder in individuals is caused by the inheritance of two alleles from the parents. This review focuses on various techniques that are used to diagnose or predict the possibility of a genetic disorder in patients. The conventional methods of prediction of genetic disorders use family histories and lifestyle factors, this approach may decrease the prediction accuracy. Therefore, introducing genetic risk score prediction based on SNP will increase the prediction accuracy and reduce the overall screening time of medical history. These predictions are done by taking a few samples of blood or sputum from the patient and sequencing the DNA to find the gene patterns. Genetic disorders can be caused by both dominant and recessive alleles. The prediction is done by finding the gene in a sequence that is increased or decreased in size; this is called Copy Number Variation (CNV). There are many studies focused on finding the correlation between the CNV of two different genomes. Researchers used many techniques to find the correlation between CNVs including machine learning, signal processing techniques. We carefully analyzed more than 50 peer-review journals and compared various methods to find the similarity in various techniques.

Index Terms: Genetic Disorder, Classification Techniques, Machine Learning, Genome Data, Copy Number Variation.

DOI: 10.47750/pnr.2022.13.S03.019

INTRODUCTION

Our modern science opened a greater understanding of genetic diseases after the discovery of genome sequencing techniques to detect mutation variants in the genome. This led to an understanding of the correlation between certain phenotypes and their corresponding similarities with genome variants. Due to the advent of parallel sequencing techniques, we can find patterns in point mutation and whole chromosomal rearrangements. This kind of comparative study shows a better insight into the type of mutations and their rate which are causing any particular genetic disorder in patients [1].

After Whole Genome Sequencing (WGS) is done, CNV can be found with the help of four different techniques read pair and read depth, split read, and assembly-based methods. Many software packages are used to read CNV in which Break Dancer and Delly found to have sensitivity above 93% [2]. Either increased or decreased the number of centrioles is the cause of genetic disorders. A most common way to group genome-based Mendelian disease is to group the genome sequence based on the phenotype and semantic similarities. There is more than 7300 Mendelian disease discovered to date [3].

Genome-Wide association is used to associate variation in SNPs with specific phenotypes in patients. This variation in SNP in the genome is called Copy Number Variants (CNV). Copy number variation is deletion, creation, and duplication of a section of the genome which constitutes a genetic difference between individuals. CNV in the genome has recurrent patterns with high mutation. By analyzing the CNV pattern across the genome we can predict the association and difference between two different factors. For example, to find whether a person with a specific disease has the probability of getting a certain phenotype, a group of people's DNA sequences is analyzed and a correlation value is used to find the similarity [4].

People with specific phycological disorders show increased rates of having other diagnoses as well. Many genes have high penetrance in inheriting certain types of cancer [5]. Various studies in genetic psychiatric disorders show variable phenotypic variation across individuals. Observations of the genetic diagnostic approach show greater authentication in biological validity. This brings us to the importance of discoveries in research strategies in the cross between dis-

eases [6]. Studies in genetic epidemiology show that people with first-degree patients with schizophrenia have an increased risk of inheriting BP as well. The heritability rate of bipolar disorder and schizophrenia is about 59% and 64%. The research shows that the risk rate of full siblings is much larger than the half-siblings. The rate of inheriting both BP and Sch is about 63% [7]. There are more than 450 SNPs that are associated with similarities between BP and Sch [8]. Our main aim is to investigate the correlation between genetic patterns and different heritable diseases.

All the existing studies to detect the manifestation or susceptibility of the disease are through family lineage with a limited genetic marker. These techniques are highly unrealistic because the influence of genes in diseases doesn't just depend on family lineage. There are many gene target strategies in which each gene that is affected by a particular disease is analyzed and compared. This type of candidate gene association study can be very limiting as it can only focus on a particular gene. This disadvantage can be eliminated by Gene Wide Association Studies (GWAS) which study the pattern of the gene throughout the genome. The basic concept behind GWAS is by studying common genetic variants occurring in each SNP of patient affected by disease and analyzing whether that same modification is measured in patients that are not affected by that disease. This will be used to help to mark a particular SNP that affects the most in genetic risk of the patient [9]. This coinheritance of specific genes more frequently for a particular effector gene is called linkage equilibrium and if the affected gene SNP is far away from the effector gene then it is called linkage disequilibrium [10].

After GWAS, with a better understanding of the SNP and its effect on particular diseases, detecting a set of a population that can pose the risk of inheriting the disease is very important. This can be done with the help of various Polygenic Risk Scoring (PRS) methods. PRS values can be calculated for all the types of diseases which can be affected by heredity. For example, PRS values for the large population of people are calculated and the people with PRS values above 90% percentile in the distribution are given special care and prescription. This can greatly increase the healthcare system's accuracy and save many lives at the same time [11]. These methods need to be analyzed extensively for a better understanding of disease diagnosis methods and also to improve the testing methods to save as many lives as possible. In this paper, we divided whole risk detection methods into various categories and discussed them according to their efficiency and accuracy, etc.

METHODS

The are many methods and ways to estimate the ways to measure the effect of genetic variants in different diseases. Some methods use the whole genome data as a sequence and find the association between disease and a gene variant. The other types use the correlation between each gene variant with the disease database to find the similarity value. Both of these techniques are credible in their ways. In this section, we are going to analyze all the types of studies related to this field.

A. Genome-Wide Association Studies (GWAS)

Considering the genetic sequence as a signal to find out the pattern is considered effective for dynamic prediction. For example, if the objective of computation is to find a specific genetic disorder pattern in patients. Catherine Stamoulis (2011) used array Comparative Genomic Hybridization (aCGH) to find the CNVs across the genome. The author used matched filtering method to match two different CNV of cancer positive and negative cancer genome collected from the ATLAS database with glioblastoma database. The proposed approach using signal decomposition methods find to have a higher true detection rate when compared to conventional methods. The important advantage of this method happens when analyzing large databases [12]. Aiello-Laws et al (2011) estimated the cancer risk in patients with the help of genetic testing, patient's family pedigree monitoring and health data monitoring, etc. The value of penetrance is calculated for mutated genes in the genome and penetrance values are used to estimate the genetic risk factor in individuals along with family pedigree evaluation [13]. The variants in the genes and phenotype are associated with a factor called Narrow sense heritability. The heritability can be calculated by calculating the variance of each SNP sequence in the array. Studies show that heritability is also affected by incomplete linkage and genotyped SNPs and reduced frequency of alleles invariants [14]. GWAS studies are conducted using scripts based on R languages in software like PLINK, PRSice, etc. All the data used in the software should be either in text format or binary format. Genotyping results are stored in the ". fam" file, bipolar disorder values are stored in the ". bed" file and the position of SNPs are stored in the ". bim" file. This data can contain some errors due to the poor quality of DNA samples during diagnosis. Therefore, essential QC steps are necessary for accurate GWAS analysis [15].

System genetics approaches are used to find novel treatment approaches for various diseases for prescribing precise medicine. However, association studies between risk factors and disease phenotypes may not be suitable for all situations. Therefore, Mendelian Randomization (MR) can also be used along with PRS to find the risk of disease. MR method analyses the effect of modifiable exposure of disease in observable environment. This can be analyzed by estimating the effect of disease inpatient in family tree-based environment studies. The effective use of GWAS studies includes both genetic and family environment studies, this will give a better understanding of the disease and medication procedures [16].

B. Quality Control in GWAS

The reason for some false detection or error in classification can be reduced by Quality Control procedures in input data. But the main disadvantage of existing GWAS pre-processing methods are highly intensive in computation and difficult. An important aspect of the Quality Control procedure is to check sample identity and pedigree integrity by analyzing sex inconsistencies and sample mix-ups in the data. The data processing is done by adding markers or annotations to the SNP data. Then the SNP with a call rate lesser than 95% is removed. Call rate is defined as the ratio of SNP in the genome which doesn't have missing data. Then there are many steps in QC which are sex check, ancestry marker check and duplicate concordance check, etc [17].

C. Polygenetic Risk Scoring (PRS)

The method of calculating the contribution of a set of SNP to a particular genetic disease is called Polygenetic Risk Scoring (PRS). There are two types of PRS weighted and unweighted. In unweighted PRS the results are found by just adding the resultant values of each of the risky alleles. Weighted PRS is done with the assumption that each different alleles affect the outcome in a very different manner, therefore, the weight of each allele is calculated along with its risk values to find the resultant PRS values [18]. Both of the processes have merits in their way. PRS values are used to access the connection between genome variations and phenotype in an individual. The basic PRS estimation was done by separating the input data into two types test data and training data. Both these data contain information about GWAS and phenotype association. The data are then processed with the help of Quality Control (QC) protocols and then the resultant data go through GWAS effect size shrinkage. Each of the GWAS data is assigned to a target population and the value of linkage disequilibrium between SNP is calculated (SNP-LP) [19].

There are certain advantages of using PRS values first one is setting up search goals to find the probability of inheriting that particular disease and the second one is specific inclusion criteria values can be added to the clinical trials to enhance the diagnosis accuracy. The final factor for calculating PRS values is they can give a basic understanding of whether that particular missing inheritance is due to misdiagnosis or some other reason. This can improve the total diagnosis process [20]. The model accuracy or data accuracy in PRS estimation can be done with the help of Area Under Curve (AUC) and Receiver Operating Curve (ROC). Basically, for the study, SNP values that have significant values above $\rho > 5\times10^{-8}$ are considered for evaluation. PRS value for breast cancer using clinical risk model was performed by **Yiwey Shieh et al (2016)** which shows an AUC value of about 0.62 [21].

D. Genetic risk assessment

The risk assessment of a particular genetic disorder is measured in factors such as penetrance, copy number variant score, etc. Risk assessment is carried out by calculating the number of allele variants that are repeated in the genome data. Jing Tan et al (2020) proposed a method to evaluate the risk of 249 autosomal recessive mitochondrial disorders using the correlation between the central database and gene data. The dataset provides alleles frequency of gnomAD database that is been compared with single-gene database using loss function estimation. This data calculated the lifetime risk of individuals from different ethnic backgrounds. The average lifetime risk is about 48-4 for 2000 samples [22]. Some studies are related to studying the correlation between neuromotor functioning and autism using genome correlation. Fadila Serdarevic et al (2019) proposed a study to find the correlation between attention-deficit/hyperactivity disorder (ADHD) children in a genome-wide association study. They used a linear regression model to find the Polygenic Risk Score (PSR) of individuals. These scores are then compared with the autistic trails scores, this method is used to find the association between neuromotor development and autism in children [23].

This brings the research to the next stage of finding the correlation between genes and human phycology. The research done by Mikaela K. Dimick et al (2019) made a study to find the association between some genetic variants and bipolar disorder in young adults. Using multigene risk score (MGRS) four gene variants are associated with bipolar disorder. The genes showing more association are IL1β rs16944, DISC1 rs821577, ZNF804A, and rs1344706. The authors also clearly predicted that this result does not predict the diagnosis of bipolar disorder in adults [24]. The research was done by N. Carmiol et al (2014) focuses on finding the genetic factors that are similar for the cause of inheritance of both alcohol use disorder (AUD) and bipolar disorder (BD). The main aim of their research is to find similar genetic variants in people that are affected by both AUD and BP. The researchers used six different phenotypes to calculate bivariate polygenic analyses between phenotypic, genetic, and environmental variables. The same test was conducted between age and sex variants in three different bivariate models. They concluded that there is a genetic correlation between alcohol use, BD, and drug use [25]. With the use of cross-referencing techniques, we can analyze the genetic factors and their effect on Bipolar patients. Some techniques use both genetic data processing and MRI image classification techniques to estimate susceptibility of BP disorder in patients [26].

E. Machine learning approaches for genetic risk assessment

Due to a large number of data human intervention can make diagnosis or analysis a little difficult. To fully utilize the

computational power of machines many machine learning algorithms are used for a better understanding of data. Genetic Algorithms (GA) are used to find the multilocal features that are present in the multiple SNP patterns. To reduce the search space in algorithm scoring for each promising gene was given. This will improve the search efficiency of the algorithm. In comparison to random search algorithms, GA was found to have higher efficiency in finding genes with lesser P-values [27]. Due to the increased understanding of genetics using Polygenic risk scores with machine learning algorithms to predict and prescribe medicine is growing in recent years. Machine learning algorithms' ability to handle multi-dimensional data increased prediction accuracy. This type of prediction increases the efficacy rate of drugs, customization of drugs, and dose usage for patients. In a study, there are about 200 data are considered to find the polygenic variance between different SNPs using gradient boosted regression tree models. They used both machine learning and genome-wide metadata analysis to find the polygenic traits in the individuals [28].

In cancer detection using polygenic risk, scores are performed by combining both SNP and genome-wide risk prediction. This can improve the accuracy of prediction in complex cancer phenotypes. Minta Thomas et al (2020) combined both loci-wise and Bayesian genome-wide association studies to find Accurate colorectal cancer (CRC) in patients. They conducted both affected case studies and control studies for patients to find the correlation in the data. But this combination can only find the top 10% of the people having similar relative 90% of people found to have no connection to study whatsoever. These patients needed further investigation down the line to predict with improved accuracy [29]. This brings out the next question are machine learning techniques effective for the prediction of novel genetic diseases. Researchers have found a tool called DOMINO to find the dominant mutation for Mendelian disorders. The tool has machine learning to extract genomic data, gene expressions, and protein interaction, and protein structures. The tool used supervised learning to predict the genomic variation and the results are found to have a 0.92 AUC value. The specificity value of the proposed method is about 87.4% and the sensitivity value is 80.4% which is better when compared to previous techniques [30].

Table 1: Machine learning approaches for cancer risk estimation

Method	Classifier	Detected disorder	AUC	Sensitivity	Specificity	Accuracy
Ali Muhamed Ali et al (2018) [31]	LSTM	Kidney Cancer	-	0.95	0.94	95%
Kaishan Tao et al (2020) [32]	Statistical model	Hepatocellular carcinoma	0.89	0.56	0.96	-
Adetiba, E., et al (2015) [33]	SVM	Lung cancer	-	-	-	95.9

After genetic disorder detection using machine learning to predict psychiatric disorder using genome data. Existing phycological disorder detection techniques use brain imaging techniques by detecting the change in regions of the image. These imaging techniques are only used for diagnostic purposes, but for predicting risk factors genome-based detection is necessary. Many studies show a major connection between inheritance and psychiatric disorders. Therefore, genome-wide association studies are of the most importance. Schizophrenia, bipolar, and autism diseases genes can be analyzed with the help of regression-based analysis in machine learning. Genetic data are separated into test and training sets and analyzed based on MADRS scores. The regression analysis methods always need a classifier in the end to predict the results. The classifiers include Naïve Bayes, KNN (K-nearest neighbor), ANN and SVM, etc... In SVM the regression data are then split into two different models and then classified based on data points. SVM can be effective for comparison between two different data set, but they are ineffective for overlapping classes in the dataset. This makes the classifier ineffective for large datasets (Malgorzata Maciukiewicz et al) [34]. Mental health can be predicted by the KNN algorithm by finding the distance between two different data. KNN algorithms can be used for both classification and regression-based problems. For beginner's genome data of test and disease data are clustered and K value or distance value from test genome data and diseases data are calculated. The final results are analyzed by finding the shortest distance between test and training data. This is an intense based training method as all the computations are done in the training phase itself. The main drawback with KNN based classification technique is it takes a lot of memory during computation and it is not suitable for large and complex datasets (Srividya, M. et al (2018)) [35]. Using a convolutional neural network to predict a pattern in genomic data is not new. Each different chromosome data is taken as input in the CNN input layer. And the models are trained using a gradient descent algorithm to find the results. CNN can also use classifiers such as decision trees and random forest algorithms. CNN algorithms classification accuracy was about 65% with an AUC of about 0.56 (Laksshman, S., et al (2017)) [36].

Table 2: Machine learning approaches for Phycological disorder risk estimation

Method	Classifier	Detected disorder	AUC	Sensitivity	Specificity	Accuracy
GBT [37]	Tree based	schizophrenia	0.95	84.9	86.6	85.7
Data mining [38]	SVM	Mood disorders	0.52	-	-	80.4
Machine learning [39]	KNN and SVM	Autism and depression	0.75	-	-	77
Machine learning [40]	Random Forrest and KNN	Bipolar Disorder	-	-	-	77.4
Murat Go¨ k et al (2018) [41]	Naïve bayes	Autism Spectrum Disorder	0.583	0.902	-	78.31

DISCUSSION

In this review, we focused on various techniques that are used to find the potential risk of genetically inherited diseases. With a greater number of techniques to find the risk of disease, there are also some problems of inconclusiveness and uncertainty in the results. Improved testing accuracy can save many lives, but the maximum sensitivity that is achieved currently doesn't go beyond.95, which is lower when u consider this as an important diagnostic method. As we discussed earlier many techniques use both GWAS, PRS, and as well as imaging and blood sampling methods to predict the results. This kind of hybrid method can increase the diagnostic accuracy for patients [42][43]. Some techniques use mendelian traits, genome data, and image data which happens to increase the diagnosis timing [44]. Some studies show that PRS methods can increase depression in children who are facing trauma from childhood [45]. Many studies focus on randomized control trials, the main disadvantage with this study is that focusing on only a particular SNP allele is ineffective. Therefore, large-scale GWAS techniques were carried out [46]. To make accurate precision medicine for complex diseases Polygenic Risk Scoring methods were used. But the main disadvantage with PRS methods as discussed above are they are unidirectional, completely independent predictor results and cannot be accountable for complex diseases [47]. These drawbacks can be eliminated with the use of machine learning models in the methods. These new models are extremely effective for large-scale complex databases and reliable for precise diagnosis. Machine learning models can be used for both cancer risk prediction as well as phycological disorder prediction methods. After analyzing all the different types of methods, it is evident that each type of disease needs a different approach for predicting the risk score. There cannot be a simple one model fits all solution for these problems. Therefore, a clear understanding of the type of input and objective of our diagnosis should be clear before choosing the correct model.

FUTURE SCOPE

The hybrid methods of using two different data samples for analysis can show better results, but the main disadvantage with this technology is that it may bring some human error. This is because both the data need different diagnostic methods to analyze them and there is no common architecture to bring the two methods together. Therefore, in the future make these hybrid systems use a universal rule and architecture has to be finalized. This can reduce the total time taken for diagnosis and also can reduce the output error. Machine learning-based models can improve the accuracy but at the same time, they also need a very large dataset as input for better training. Even after the prediction, there are various issues with the interpretation of output data. These disadvantages need to be addressed and rectified in future research. Association of genetic risk score results and depression in patients cannot be ignored easily. There must be some guidelines, structure, and ethical aspects of it also need to be framed.

CONCLUSION

After analyzing various papers on genomic association studies, it is evident that accurate prediction of different variants is the need of the hour. The main problem that arises with these different methods is there is no universal framework or roadmap that is present for risk prediction models. This can greatly reduce the inconsistency and accuracy problems in various methods. Analyzing different methods makes it clear that there need to be various prediction systems in place because of data inconsistency issues. To achieve the highest efficiency genetic studies needs to be linked with phenotyping studies. Improvement in these methods can give a better understanding of the etiology of various genetically caused diseases, which will lead to better accuracy in medical treatments.

REFERENCES

Goswami, R. S., & Harada, S. (2020). An Overview of Molecular Genetic Diagnosis Techniques. Current protocols in human genetics, 105(1), e97.

Whitford, W., Lehnert, K., Snell, R. G., & Jacobsen, J. C. (2019). Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. Journal of biomedical informatics, 94, 103174.

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., & Robinson, P. N. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated ge-

nome. Science translational medicine, 6(252), 252ra123-252ra123.

Alqallaf, A. K., Tewfik, A. H., Krakowiak, P., Tassone, F., Davis, R., Hansen, R., & Selleck, S. B. (2009, May). Identifying patterns of copy number variants in case-control studies of human genetic disorders. In 2009 IEEE International Workshop on Genomic Signal Processing and Statistics (pp. 1-4). IEEE.

French, J. D., & Edwards, S. L. (2020). Genetic determinants of breast cancer risk. Current Opinion in Endocrine and Metabolic Research.

Finucane, B. M., Ledbetter, D. H., & Vorstman, J. A. (2021). Diagnostic genetic testing for neurodevelopmental psychiatric disorders: Closing the gap between recommendation and clinical implementation. Current opinion in genetics & development, 68, 1-8.

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet, 373(9659), 234-239.

Purcell, S. (2009). International Schizophrenia ConsortiumWray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460, 748-752.

Guo, X., & Rotter, J. I. (2019). Genome-wide association studies. Jama, 322(17), 1705-1706.

Lange, K. E. N. N. E. T. H. (1993). A stochastic model for genetic linkage equilibrium. Theoretical population biology, 44(2), 129-148.

Francisco, M., & Bustamante, C. D. (2018). Polygenic risk scores: a biased prediction?. Genome medicine, 10(1), 1-3.

Stamoulis, C., & Betensky, R. A. (2011). A novel signal processing approach for the detection of copy number variations in the human genome. Bioinformatics, 27(17), 2338-2345.

Aiello-Laws, L. (2011, February). Genetic cancer risk assessment. In Seminars in oncology nursing (Vol. 27, No. 1, pp. 13-20). WB Saunders.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics, 42(7), 565-569.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research, 27(2), e1608.

Erdmann, J., Kessler, T., Munoz Venegas, L., & Schunkert, H. (2018). A decade of genome-wide association studies for coronary artery disease: the challenges ahead. Cardiovascular research, 114(9), 1241-1257.

Highland, H. M., Avery, C. L., Duan, Q., Li, Y., & Harris, K. M. (2018). Quality control analysis of Add Health GWAS data. Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: polygenic risk score software. Bioinformatics, 31(9), 1466-1468.

Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2018). A guide to performing Polygenic Risk Score analyses. BioRxiv, 416545.

Escott-Price, V., Myers, A. J., Huentelman, M., & Hardy, J. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Annals of neurology, 82(2), 311-314.

Shieh, Y., Hu, D., Ma, L., Huntsman, S., Gard, C. C., Leung, J. W.,... & Ziv, E. (2016). Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast cancer research and treatment, 159(3), 513-525.

Tan, J., Wagner, M., Stenton, S. L., Strom, T. M., Wortmann, S. B., Prokisch, H.,... & Klopstock, T. (2020). Lifetime risk of autosomal recessive mitochondrial disorders calculated from genetic databases. EBioMedicine, 54, 102730.

Serdarevic, F., Tiemeier, H., Jansen, P. R., Alemany, S., Xerxa, Y., Neumann, A.,... & Ghassabian, A. (2020). Polygenic risk scores for developmental disorders, neuromotor functioning during infancy, and autistic traits in childhood. Biological psychiatry, 87(2), 132-138.

Dimick, M. K., Cazes, J., Fiksenbaum, L. M., Zai, C. C., Tampakeras, M., Freeman, N.,... & Goldstein, B. I. (2020). Proof-of-concept study of a multi-gene risk score in adolescent bipolar disorder. Journal of affective disorders, 262, 211-222.

Carmiol, N., Peralta, J. M., Almasy, L., Contreras, J., Pacheco, A., Escamilla, M. A.,... & Glahn, D. C. (2014). Shared genetic factors influence risk for bipolar disorder and alcohol use disorders. European Psychiatry, 29(5), 282-287.

Sarıçiçek, A., Yalın, N., Hıdıroğlu, C., Çavuşoğlu, B., Taş, C., Ceylan, D.,... & Özerdem, A. (2015). Neuroanatomical correlates of genetic risk for bipolar disorder: a voxel-based morphometry study in bipolar type I patients and healthy first-degree relatives. Journal of affective disorders, 186, 110-118.

Mooney, M., Wilmot, B., & McWeeney, S. (2011). The GA and the GWAS: using genetic algorithms to search for multilocus associations. IEEE/ACM transactions on computational biology and bioinformatics, 9(3), 899-910.

Paré, G., Mao, S., & Deng, W. Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. Scientific reports, 7(1), 1-11.

Thomas, M., Sakoda, L. C., Hoffmeister, M., Rosenthal, E. A., Lee, J. K., van Duijnhoven, F. J.,... & Hsu, L. (2020). Genome-wide modeling of polygenic risk score in colorectal cancer risk. The American Journal of Human Genetics, 107(3), 432-444.

Quinodoz, M., Royer-Bertrand, B., Cisarova, K., Di Gioia, S. A., Superti-Furga, A., & Rivolta, C. (2017). DOMINO: using machine learning to predict genes associated with dominant disorders. The American Journal of Human Genetics, 101(4), 623-629.

Muhamed Ali, A., Zhuang, H., Ibrahim, A., Rehman, O., Huang, M., & Wu, A. (2018). A machine learning approach for the classification of kidney cancer subtypes using mirna genome data. Applied Sciences, 8(12), 2422.

Tao, K., Bian, Z., Zhang, Q., Guo, X., Yin, C., Wang, Y.,... & Xing, J. (2020). Machine learning-based genome-wide interrogation of somatic copy number aberrations in circulating tumor DNA for early detection of hepatocellular carcinoma. EBioMedicine, 56, 102811.

Adetiba, E., & Olugbara, O. O. (2015). Lung cancer prediction using neural network ensemble with histogram of oriented gradient genomic features. The Scientific World Journal, 2015.

Maciukiewicz, M., Marshe, V. S., Hauschild, A. C., Foster, J. A., Rotzinger, S., Kennedy, J. L.,... & Geraci, J. (2018). GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. Journal of psychiatric research, 99, 62-68.

Srividya, M., Mohanavalli, S., & Bhalaji, N. (2018). Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems, 42(5), 1-12.

Laksshman, S., Bhat, R. R., Viswanath, V., & Li, X. (2017). DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. Human mutation, 38(9), 1217-1224.

Trakadis, Y. J., Sardaar, S., Chen, A., Fulginiti, V., & Krishnan, A. (2019). Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 180(2), 103-112.

Pirooznia, M., Fayaz Seifuddin, J. J., Mahon, P. B., Potash, J. B., Zandi, P. P., & Bipolar Genome Study (BiGS) Consortium. (2012). Data mining approaches for genome-wide association of mood disorders. Psychiatric genetics, 22(2), 55.

Cho, G., Yim, J., Choi, Y., Ko, J., & Lee, S. H. (2019). Review of machine learning algorithms for diagnosing mental illness. Psychiatry investigation, 16(4), 262.

Acikel, C., Son, Y. A., Celik, C., & Gul, H. (2016). Evaluation of potential novel variations and their interactions related to bipolar disorders: analysis of genome-wide association study data. Neuropsychiatric

disease and treatment, 12, 2997.

Gök, M. (2019). A novel machine learning model to predict autism spectrum disorders risk gene. Neural Computing and Applications, 31(10), 6711-6717.

Xu, Z., Wu, C., Pan, W., & Alzheimer's Disease Neuroimaging Initiative. (2017). Imaging-wide association study: Integrating imaging endophenotypes in GWAS. Neuroimage, 159, 159-169.

Knutson, K. A., Deng, Y., & Pan, W. (2020). Implicating causal brain imaging endophenotypes in Alzheimer's disease using multivariable IWAS and GWAS summary data. NeuroImage, 223, 117347.

Zhou, Y., Srinivasan, S., Mirnezami, S. V., Kusmec, A., Fu, Q., Attigala, L.,... & Schnable, P. S. (2019). Semiautomated feature extraction from RGB images for sorghum panicle architecture GWAS. Plant Physiology, 179(1), 24-37.

Peyrot, W. J., Milaneschi, Y., Abdellaoui, A., Sullivan, P. F., Hottenga, J. J., Boomsma, D. I., & Penninx, B. W. (2014). Effect of polygenic risk scores on depression in childhood trauma. The British Journal of Psychiatry, 205(2), 113-119.

XU, J., XIA, J., & ZHENG, H. (2018). Assessing the clinical risk factors of fragility fractures with GWAS data and Mendelian randomisation approach. Chinese Journal of Endocrinology and Metabolism, 987-991.

Leonenko, G., Sims, R., Shoai, M., Frizzati, A., Bossù, P., Spalletta, G.,... & Escott-Price, V. (2019). Polygenic risk and hazard scores for Alzheimer's disease prediction. Annals of clinical and translational neurology, 6(3), 456-465.

Ibrahim, S., & Koksal, M. E. (2021). Commutativity of sixth-order time-varying linear systems. Circuits, Systems, and Signal Processing, 40(10), 4799-4832.