Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 1758-1773
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.2338

Machine learning-driven polygenic risk scores for bipolar disorder, depression, and panic disorder

Sara Benoumhani[1,2], Saima Jabeen[1,2], Mariam M AlEissa[1,3,4*]
[1]AI Research Center, College of Engineering, Alfaisal University, Riyadh, Saudi Arabia; sbenoumhani@alfaisal.edu (S.B.)
[2]Software Engineering Department, College of Engineering, Alfaisal University, Riyadh, Saudi Arabia; sjabeen@alfaisal.edu (S.J.)
[3]Molecular Genetics Laboratory, Public Health Authority, Riyadh, Saudi Arabia; mariam_al_eissa@hotmail.com (M.M.A.E)
[4]College of Medicine, Alfaisal University, Riyadh, Saudi Arabia.

Abstract: Polygenic Risk Score (PRS) is a computational tech- nique that uses various genomic data to simultaneously analyze an individ-ual's genetic risk for particular illnesses or traits. However, the traditional PRS computation has a few weaknesses, including its limited capacity to account for just a portion of trait variance, susceptibility to overfitting, and insufficient ability to discriminate among the larger population. Machine Learning (ML) methods offer a promising alternative to the traditional method by avoiding the problem of overfitting and improving accuracy. This study aims to develop an ML model for improved PRS calculation. We used the summary statistics for three mentals diseases, bipolar, depression, and panic disorder, from the Psychiatric Genomics Consortium (PGC) as a disease reference. We also obtained actual genotype data of individuals from OpenSNP, which includes both case and control samples. This data is used for predicting scores. The suggested approach, called Polygenic Risk Score Neural Network (PRSNN), calculates the PRS using weight vectors that estimate the relevance of each single nucleotide polymorphism (SNP) with a particular phenotype by deep learning model as an alternative to the traditional method. This study aims to develop a machine learning model, called PRSNN, for improved calculation of Polygenic Risk Scores (PRS). The PRSNN method outperforms the conventional method in identifying individuals at risk of mental disease. A novel deep-learning approach, named as PRSNN, is proposed for generating PRSs. The results demonstrate that it outperforms the traditional method of computing PRS for complex diseases. Further upgrades for this tool are required to overcome the current limitations, including lack of validation with external data from different ancestries, which may limit the applicability of the PRSNN method across diverse populations, and the small sample size, which may affect the results.

Keywords: *Genome-wide association studies (GWAS), Machine Learning (ML), Polygenic risk score (PRS), Psychiatric genomics con-sortium (PGC).*

1. Introduction

A key function of public health is to prevent the community from diseases and provide an early intervention that measures for a specific disease based on their risk level, to pick up future cases in the early stage by implementing a targeted screening program, and to intervene before disease formation [1].

Psychiatric and neurodegenerative disorders cause substantial burdens on individuals, their families and the community, causing a high rate of morbidity and mortality [2][3][4]. Those diseases have an ongoing course and typically start in childhood or young adulthood, meaning many years are spent with a crippling condition. Furthermore, people with serious mental illness frequently have lower socioeconomic standing [5]. Other than crippling the patient's life, they regularly encounter stigma. Psychiatric disorders have higher rates of both drug use and alcohol usage, all of which have a detrimental effect on one's healthand quality of life [6] [7][8]. Estimates

* Correspondence: mariam_al_eissa@hotmail.com

indicate that individuals grappling with serious mental illness have an average life expectancy approximately ten years shorter than that observed in the general population [9] [10].

To study how genes influence complex traits or diseases and to predict who is more likely to have them, researchers use simple models that add up the effects of many genetic variants. These models are called Polygenic Risk Scores (PRSs) [11]. The traditional method to generate PRS is the most common method to calculate PRSs by summing the number of SNP a person has multiplied by the weight of each SNP [12]. The effect weight of the SNP on the trait or disease is represented by the weight, estimated from the previous Genome-wide Association Studies (GWAS) [13], which are techniques that identify associations between phenotypes and genotypes that are involved in causing a certain disease [14]. Based on comparing cases with controls will assess the identification of SNPs that might contribute in the disease formation [15][16].

However, this assumption may not hold for complex diseases incorporated in gene-gene interactions. Therefore, some researchers have suggested using non-additive models or methods that can account for these interactions in PRS calculation [17]. Examples of non-additive models or methods include quadratic, logistic, or neural network models, as well as Bayesian methods or machine learning methods [18] [19].

In addition, the traditional method of calculating PRS has some limitations, such as requiring large training and testing datasets, the lack of generalizability across different populations, and the inability to capture the interactions and dependencies among genetic variants. Moreover, the PRS models may be biased by the ancestry of the discovery dataset and may not generalize well to other populations [20][21] [22]

It is noted that these models exhibit a restricted capacity for model fitting, thereby influencing the precision of the ultimate prediction outcomes to a certain extent. Nowadays, machine learning has contributed towards advancing various fields, including genomics with deep-learning techniques, medical imaging, and natural language processing for health records [23][24][25].

Some researchers have proposed to use machine learning methods, which can handle high-dimensional and complicated data and learn from patterns and characteristics. ML methods can overcome some of the limitations of the traditional method of calculating PRS. For example, ML methods can rank the SNPs according to their importance or relevance for the disease risk, or theycan identify the pathways or biological processes that are associated with the disease. They can also visualize the relationships between the features and the outcome, or between different features, using graphs or networks [26].

This paper presents a deep learning-based approach, named as PRSNN, to generate polygenic risk scores (PRSs). Section 2 discusses the adopted materials and methods to explain the methodology of the proposed approach. Section 3 presents the obtained results. Discussion is made in Section 4 while Section 5 draws the conclusion.

2. Materials and Methods

We propose a methodology for disease diagnosis using genetic data. Our method- ology consists of two parts: generating polygenic risk scores and evaluating their diagnostic accuracy. We use two methods to generate PRSs: the traditional method, which selects single nucleotide polymorphisms based on their quality and effect size, and the PRSNN method, which trains a neural network model to learn the SNP weights. Our neural network model takes the SNPs, the genotype, the p-value, and the effect weights as inputs, and predicts the disease status as output. We use grid search and cross-validation to optimize the model param- eters. Once the model is trained, we use its weights to compute the PRSs by PRSNN. We test our methodology on a real-world dataset from OpenSNP and compare the diagnostic performance of the traditional method and the PRSNN method. Figure-1 illustrates the proposed methodology.

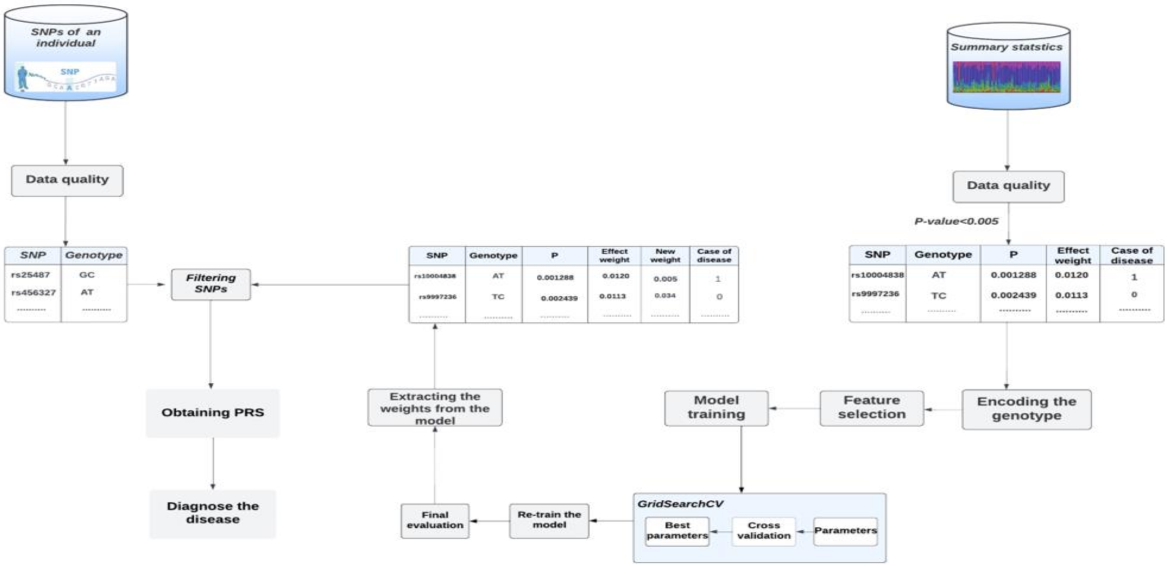
SNPs of an individual
SNP
Summary statstics
Data quality
Data quality
P-value<0.005
SNP | Genotype
rs25487 | GC
rs456327 | AT
Filtering SNPs
SNP | Genotype | P | Effect weight | New weight | Case of disease
rs10004838 | AT | 0.001288 | 0.0120 | 0.005 | 1
rs9997236 | TC | 0.002439 | 0.0113 | 0.034 | 0
SNP | Genotype | P | Effect weight | Case of disease
rs10004838 | AT | 0.001288 | 0.0120 | 1
rs9997236 | TC | 0.002439 | 0.0113 | 0
Obtaining PRS
Diagnose the disease
Extracting the weights from the model
Model training
Feature selection
Encoding the genotype
GridSearchCV
Best parameters
Cross validation
Parameters
Final evaluation
Re-train the model

Figure 1.
The proposed methodology.

2.1. Data Collection and Preprocessing

The dataset from the Psychiatric Genomics Consortium (PGC) is used, which is a large genetic variation source associated with psychiatric diseases utilizing GWAS and other methods [27][28]. We also used genotype datasets containing case and control samples from the OpenSNP platform, a website that collects personal genomic data from individuals who have done direct-to-consumer DNA testing. These genomic data contain the SNPs of the individuals, and these datasets comprise both cases of anxiety and depression as a mental disorder and healthy controls. Table 1 presents the demographic features of the OpenSNP datasets, while Figure-2 illustrates the distribution of file sources, with 23andMe, Ancestry, and FTDNA-Illumina being the most prevalent file types [29][30].

Table 1.
Demographic features of the open SNP dataset.

Demographic feature	**Details**
Age	Varies widely; includes users from different age groups
Sex	Both male and female users are represented
Ethnicity	Diverse ethnic backgrounds, particularly European
Genetic testing services	Data from multiple genetic testing services like 23andMe, deCODEme, and FamilyTreeDNA
Phenotypic traits	Users share a wide range of phenotypic traits, including physical characteristics, health conditions, and lifestyle factors
Genotypes	Over 215 million genotypes distributed across more than 2 million unique SNPs
Geographic location	Users from various countries around the world

As a result, three datasets were employed for training: the PGC summary statistics for bipolar disorder, depression, and panic disorder. For the test data, we employed the OpenSNP individual genotypes, which included traits related to anxiety, depression, bipolar disorder, and panic. The PCG library was used to detect variants in the OpenSNP dataset. Table 2 provides a summary of thefeatures of the datasets used in this study.

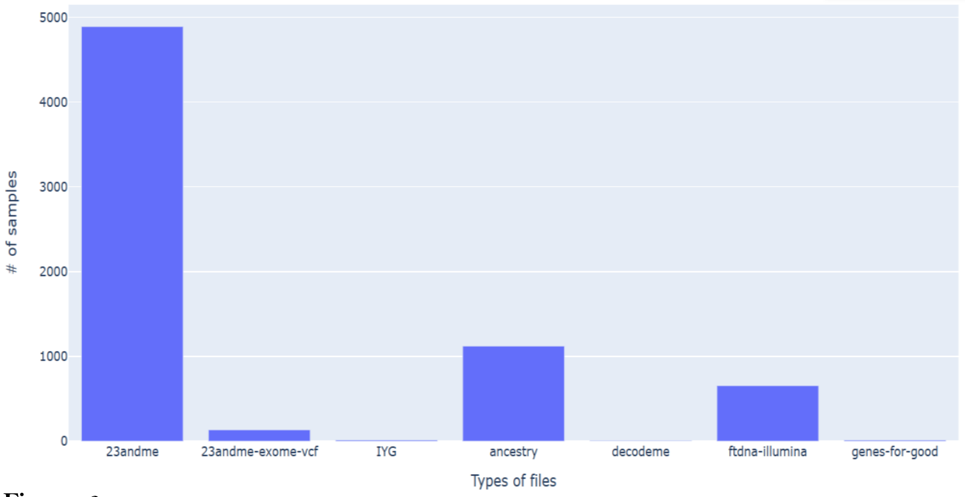

Figure 2.
Number of genotype samples available on openSNP, categorized byfile type.

Table 2.
Disease training and testing data information.

Disease	Training data	References	Testing data
Bipolar	41917 cases, 371549 controls	[31]	5 cases, 70 controls
Depression	153458 cases, 344901 controls	[32]	50 cases, 70 controls
Panic disorder	7016 cases, 14 745 controls	[33]	10 cases, 40 controls

We performed data quality checks on both sources to detect and handle ambiguous SNPs, mismatching SNPs, and missing values. We then selected a subset of SNPs from the PGC data that exhibited associations with the disease of interest, applying a p-value threshold of 0.005 to distinguish between cases and controls is considered statistically significant, particularly in the field of genomics [34]. For feature extraction, we encoded the genotype data, and for feature selection, we employed Random Forest (RF), a machine learning method, to identify the most essential features. Figure-1 shows the flowchart of our data preprocessing steps, along with some examples of SNP data and the features that we used for our model.

2.2. Deep Learning Model

A deep learning model called the sequential model, which is a kind of Neural Network (NN), is employed to predict the disease status using four features: genotype encoding, SNP, effect weight, and P-value. Our model architecture is designed to predict binary outcomes, such as disease presence, based on input data.

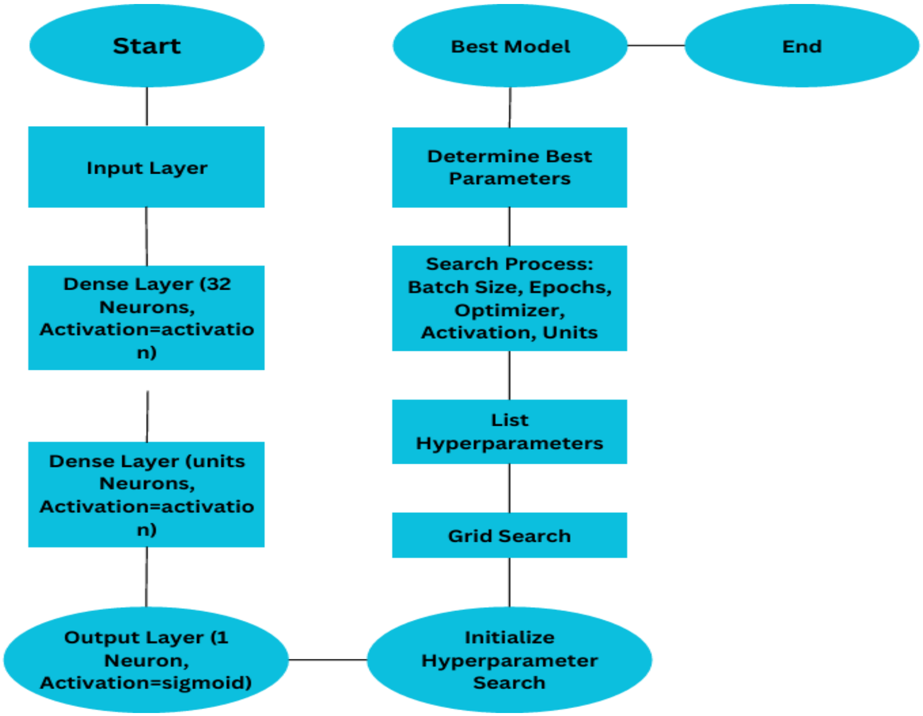
Start
Input Layer
Dense Layer (32 Neurons, Activation=activatio n)
Dense Layer (units Neurons, Activation=activatio n)
Output Layer (1 Neuron, Activation=sigmoid)
Initialize Hyperparameter Search
Grid Search
List Hyperparameters
Search Process: Batch Size, Epochs, Optimizer, Activation, Units
Determine Best Parameters
Best Model
End

Figure 3.
Flowchart of the training model and hyperparameters.

The neural network comprises two hidden layers with ReLU activationand an output layer with a sigmoid activation function. During training, we use binary cross-entropy loss, and for evaluation, we rely on accuracy. The sigmoid function produces a value between 0 and 1, representing the probabilityof disease presence. Table 3 displays the model architecture details.

Table 3.
The NN architecture details.

Layer	**Parameters**
Layer 1	Input shape
Layer 2	32 neurons
Layer 3	Units neurons (16 or 32 based on hyperparameter search) Activation = activation
Layer 4	1 neuron activation = sigmoid

To avoid overfitting and enhance the predictive accuracy of our model, we implemented cross-validation and hyperparameter tuning in our methodology. We employed the GridSearchCV module [35] to determine the optimal parameters for our model, as illustrated in Figure 3 and the prediction performance of the optimal model (the accuracy) of is equal to 0.98. The objective of our model's training is to derive the SHapley Additive exPlanations (SHAP) values, which assess the influence of each feature on the model's prediction for a given instance. They rely on a game-theoretic approach that assigns each feature a fair and consistent importance score, considering the feature interactions. The strength or intensity of the SHAP value indicates the effect. SHAP values are model-agnostic, they can provide interpretable explanations for any machine learning model, including decision trees, random forests, linear regression, and neural networks [36]. These SHAP values are the new weights, as shown in Figure 1.

2.3. Methods To Generate PRS

2.3.1. The Traditional Method

The basic stepwise process for calculating PRS adds up the effects of many small genetic variations that are linked to the trait or disease [37]. Each variation hasa weight that shows how much it influences the trait or disease. The formula to calculate the PRS for an individual is:

$$PRS_j = \sum_{i=1}^{N} \beta_i * \text{dosage}_{ij}$$

where N represents the count of SNPs in the score, β_i is the effect size (or beta) of variant i and dosage refers to the number of copies of SNP i present in the genotype of individual j [37].

2.3.2. *Polygenic Risk Score Neural Network (PRSNN)*

Some studies have used machine learning to generate PRS, such as the one by [38], which also employed weight vectors for a new PRS prediction method.However, we added more steps, such as the feature selection and modified the weight vectors obtained by the neural network. Specifically, NNW represents the weight vector derived from the NN model, as explained in the previous section. This process yields individual risk scores for each SNP. The following formula depicts this procedure:

$$PRSNN = \sum_{i=1}^{N} \text{NNW}_i * \text{dosage}_{ij} \quad (1)$$

Where the variable 'N' signifies the overlap between the SNP sets of the individual genotype and the PGC dataset. NNW vectors represent the weight vectors that assess the importance of each SNP. The dosage calculation relies on the minor allele of SNP j in the PGC data summary statistics.

2.4. *Performances*

We used logistic regression with the PRS values obtained from the traditionaland PRSNN methods to evaluate their performance in disease prediction. Wemeasured the logistic regression model using the accuracy, recall, precision, and f1 score metrics. Table 4 displays the performances.

Table 4.
Comparing traditional and PRSNN method performances in disease prediction.

Disease	Traditional method				PRSNN method			
	Accuracy	Precision	Recall	F Score	Accuracy	Precision	Recall	F score
Depression	0.8	0.82	0.97	0.89	0.96	0.97	0.97	0.97
Panic disorder	0.85	0.89	0.8	0.84	0.95	1	0.89	0.94
Bipolar	0.88	0.85	1	0,92	0.95	0.96	0.96	0.96

2.5. *Logistic Regression*

We applied logistic regression, a machine learning technique that can estimate the probability of an outcome (e.g., disease status) from one or more predictors (e.g., PRS), to predict the risk of a certain disease from genetic data. We used PRS as the sole predictor. We trained the logistic regression model on a dataset with two columns: PRS, the PRS values for each individual, and status, a binary variable indicating whether the individual had the disease or not. The model returned a coefficient for the PRS predictor that reflected the magnitude and direction of its association with the disease outcome. A positive coefficient indicated that a higher PRS increased the likelihood of having the disease, while a negative coefficient indicated the opposite. We validated the model on a new set of individuals with PRS values and used it to predict their disease status.

Figures 4, 5, 6, 7, 8, and 9 compare the actual and predicted disease diagnoses for 23 samples, using both the traditional method and the PRSNN method.

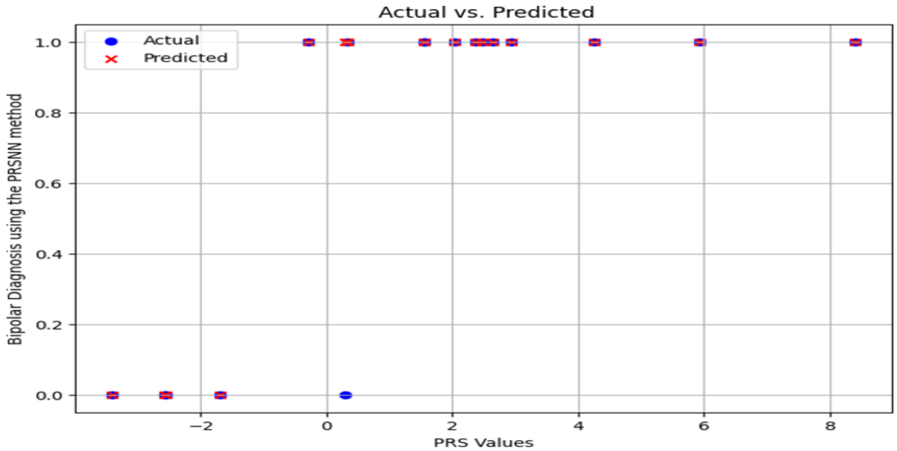

Figure 4.
Comparison of actual and predicted bipolar diagnosis using PRSNNMethod.

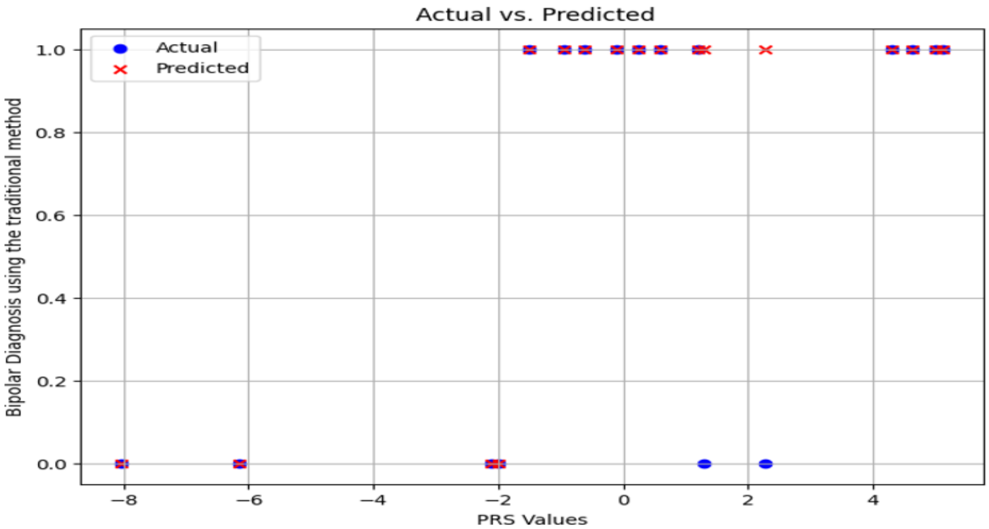

Figure 5.
Comparison of actual and predicted bipolar diagnosis using the traditional method.

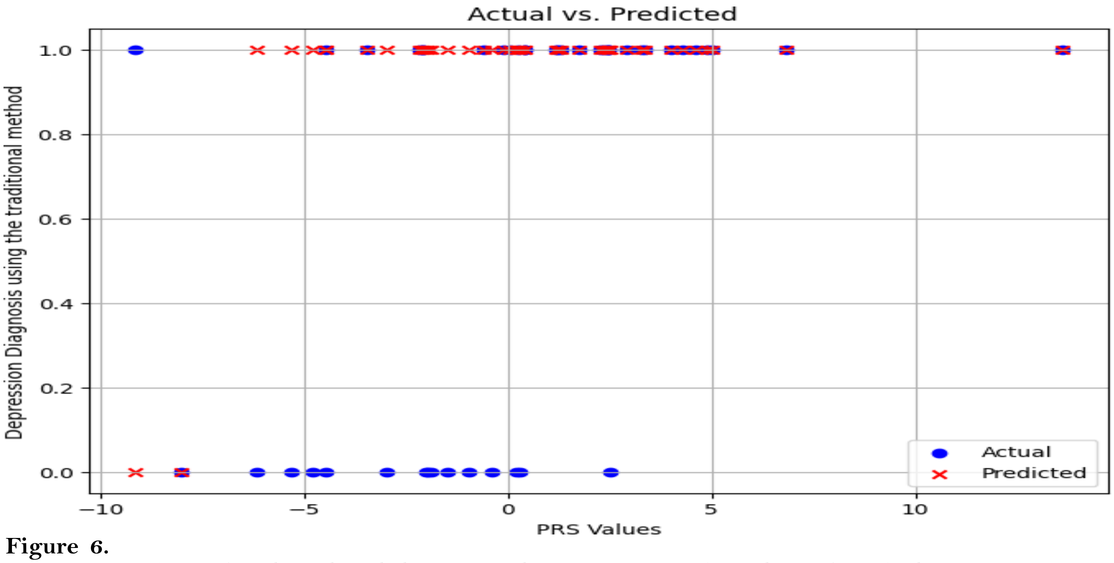

Figure 6.
Comparison of actual and predicted depression diagnosis Using thetraditional method.

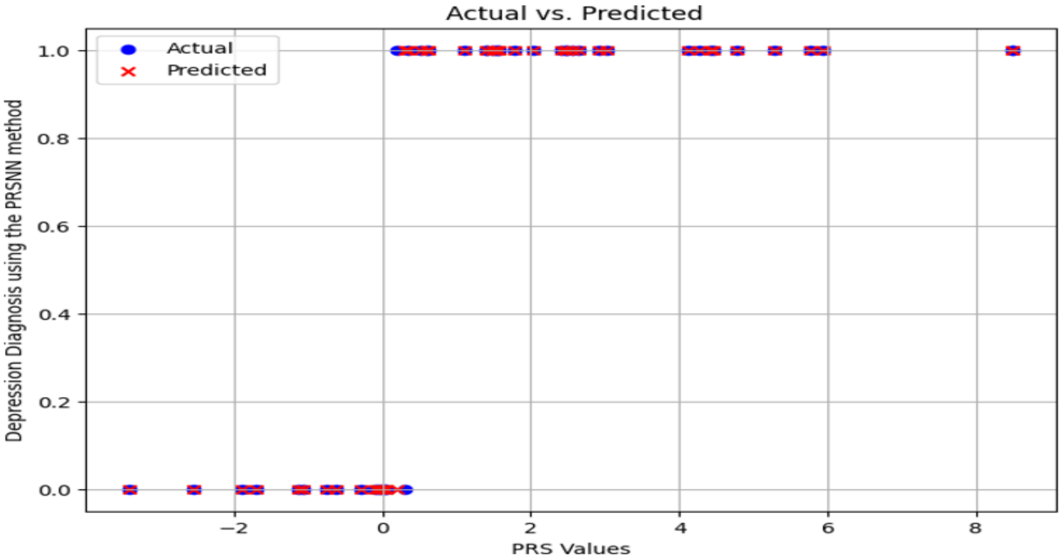

Figure 7.
Comparison of actual and predicted depression diagnosis using PRSNN method.

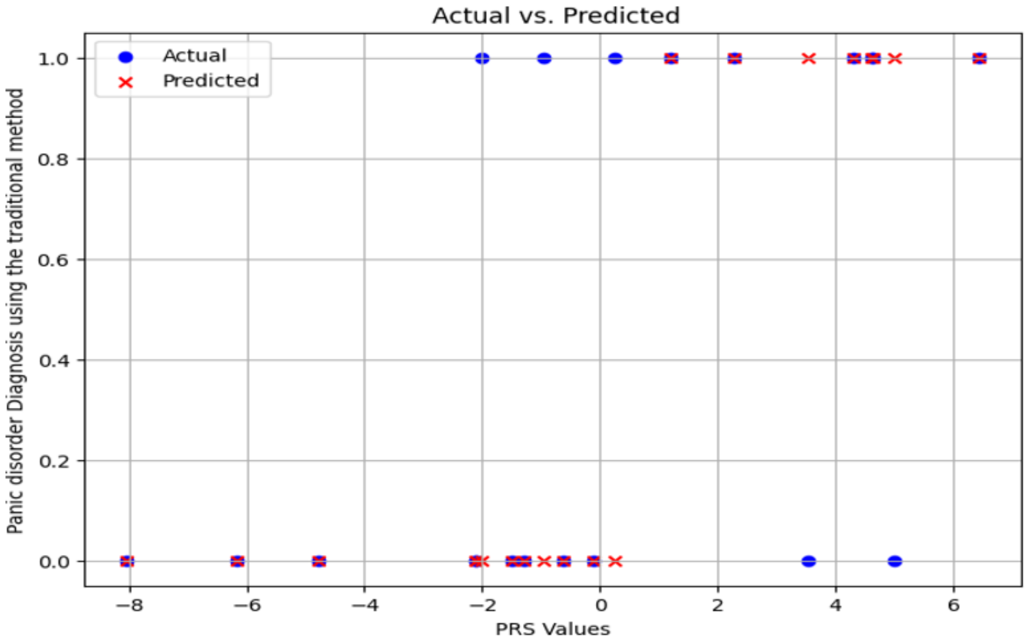

Figure 8.
Comparison of actual and predicted panic disorder diagnosis using thetraditional method.

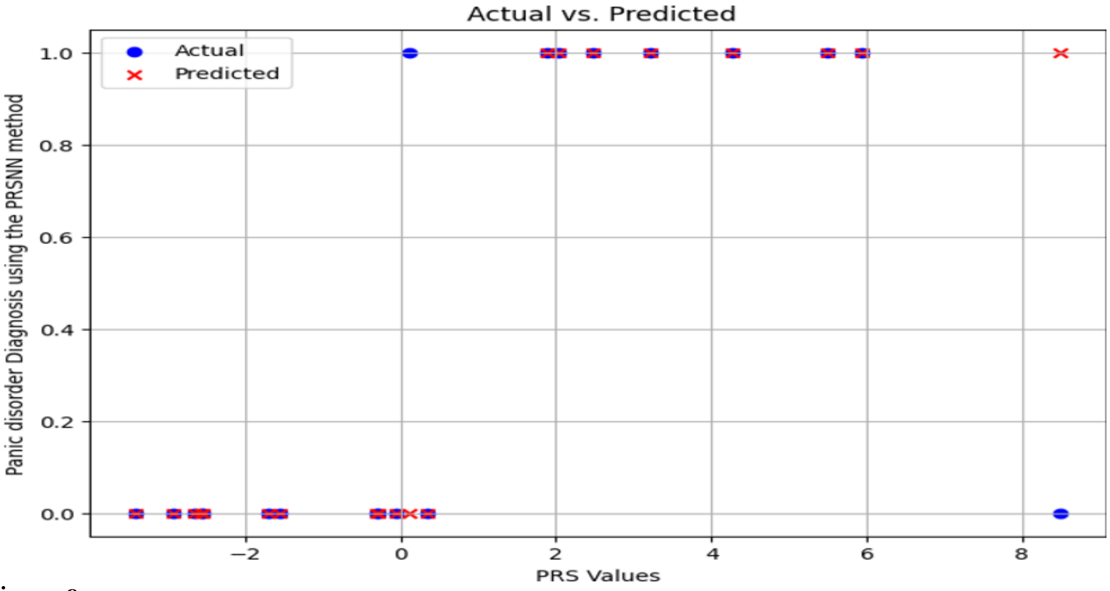

Figure 9.
Comparison of actual and predicted panic disorder diagnosis using the traditional method.

3. Results

3.1. The Distribution of PRSs Values

The box plots in Figure-10 shows the PRS values calculated by the traditional method and PRSNN respectively, show how the PRS values are distributed for people with and without bipolar disease. PRS values measure the genetic risk for these mental conditions. It is shown that people with bipolar disease have higher PRS values than people without any mental disease, which means theyhave more genetic variants linked to these conditions.

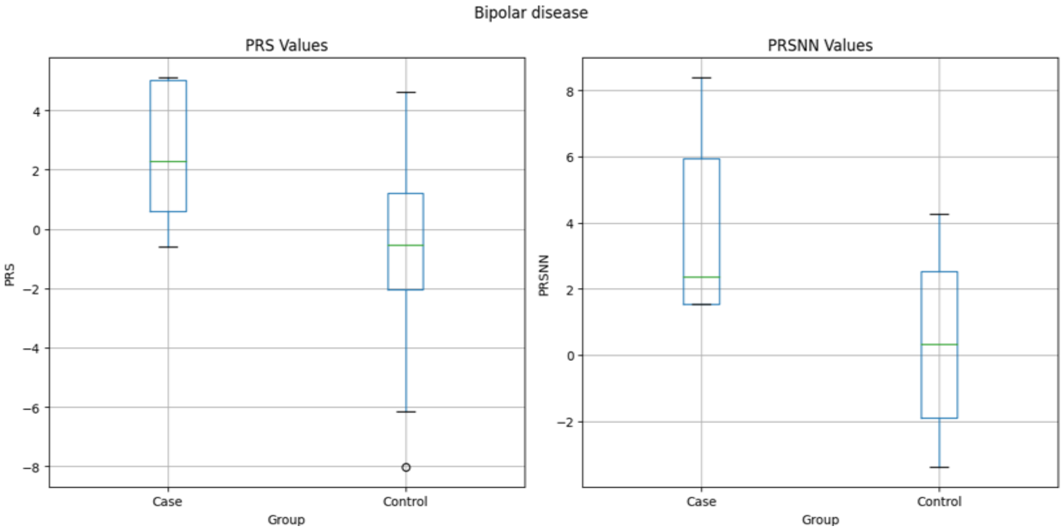

Figure 10.
Box plot case-control comparison of PRSs values compared to PRSNN values for bipolar disease.

The box plot consists of four parts: the median, the quartiles, the whiskers, and the outliers. The median is the PRS value that splits the data into two equal parts. The quartiles are the PRS values that mark the 25th and 75th percentiles of the data, which show how spread out the data are. The whiskers are the lines that go from the quartiles to the lowest and highest PRS values, except for theoutliers. Outliers are the PRS values that are much higher or lower than the rest of the data, and they are shown as circles on the box plot. The box plots in Figure-10 shows the same insights as of Figure-11 and Figure-12 but the PRSs values are calculated here for depression disease and panic disorder respectively.

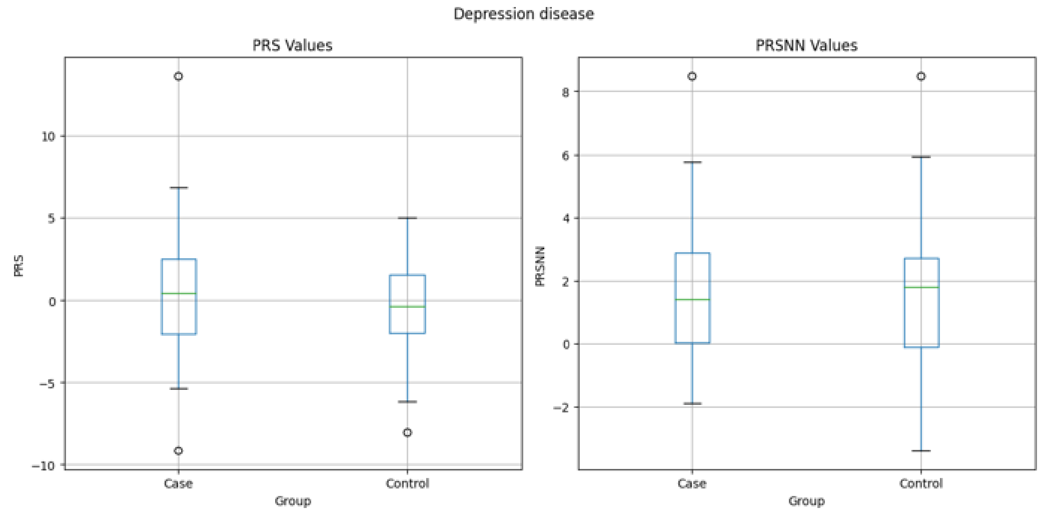

Figure 11.
Box plot case-control comparison of PRSs values compared to PRSNNvalues for depression.

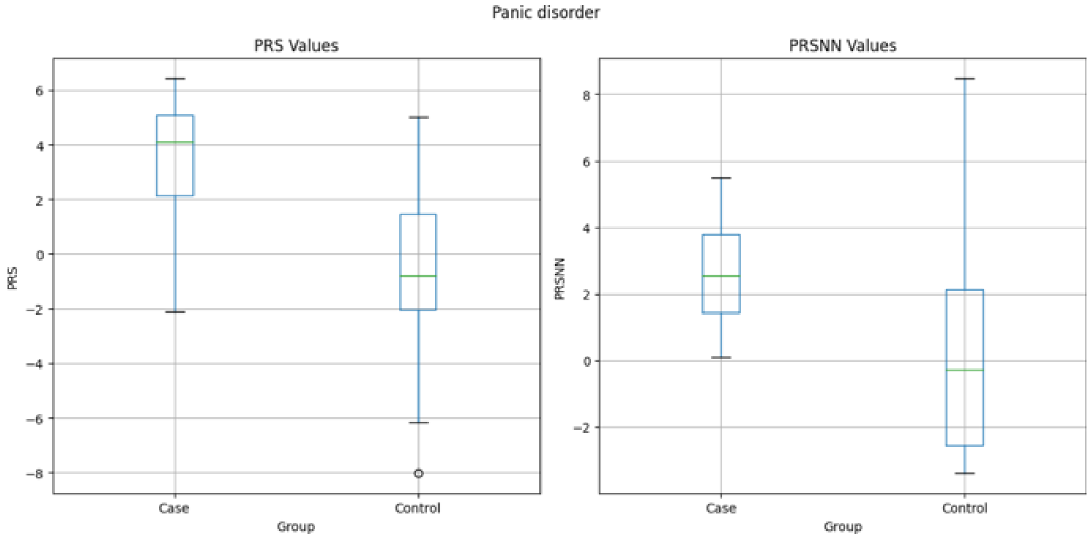

Figure 12.
Box plot case-control comparison of PRSs values compared to PRSNN values for panic disorder.

3.2. Diagnosing the Disease

3.2.1. Quartile-Based PRS Diagnosis

The box plot allows us to compare the PRS values of the two groups of people.For example, we can observe that the group with depression has a higher median PRS value than the group without any mental disease, which implies that the group with mental disease has a higher mean genetic risk. We can also notice that the group with depression has a wider interquartile range than the group without any mental disease, which implies that the group with mental disease has more diversity in their genetic risk. The whiskers and the outliers also show us the lowest and highest PRS values for each group. For example, to evaluate therisk of bipolar disease, we used the PRS distribution's quartiles. The traditiona method produces these PRS values for the 25th, 50th, and 75th percentiles: people with the disease have 0.7, 2.29, and 4.8, while people without the disease have -1, -0.3, and 1.56 (see Figure-10). Based on these values, we can diagnose the disease like this: a person with a PRS value below -1 is probably healthy, and a person with a PRS value above 0.7 is probably sick.

3.3. Comparative Analysis

[39] presented a novel deep-learning-based model named DeepPRS for scoring the risk of common diseases with genome-wide genotype data. The model was evaluated on the UK Biobank dataset and shown to outperform other state- of-the-art methods for Alzheimer's disease, inflammatory bowel disease, type 2 diabetes, and breast cancer. DeepPRS is reported not to rely solely on the additive effect of risk SNPs and has the potential to identify high-risk individuals even with few known risk SNPs.

[40] compared the performance of machine learning (ML) and deep learning (DL) methods to classical PRS calculation methods for predicting polygenic risk scores (PRS) in genome-wide association studies (GWAS). The authors used simulated GWAS data with different allele frequencies and sample sizes to evaluate the performance of different methods. They found that ML methods, such as Support Vector Machine (SVM) and Random Forest (RF), can achieve more consistent results in terms of case-control separation compared to PRS calculated with the classical method. The paper suggested that ML and DL methods can be a good alternative to classical PRS calculation methods.

[41] compared the use of PRS and machine learning for case/control classifi- cation using simulated data. The results showed that the average classification AUC for PRSice, Plink, Lassosum, and machine learning was 0.27, 0.3, 0.35, and 0.87, respectively. The authors concluded that the choice between PRS and machine learning depends on the specific phenotype. The study provides scripts and code segments for dataset generation, PRS calculation, and machine learning model training, discussing the steps involved in dataset generation, dataset division, quality control, PRS calculation, and machine learning. The PRS scores are normalized between 0 and 1, and individuals with PRS ¿ 0.5 are considered cases. While this procedure is demonstrated using simulated binary phenotypes, it can likely be adapted to real-world datasets containing continuous phenotypes by modifying the final step and employing a multi-label classification approach when training the machine learning model.

[42] also compared PRS and machine learning methods for predicting CAD status, but uses a different dataset and methodology. The authors used a dataset of 527 CAD cases and 473 controls from Germany to evaluate the performance of the same machine learning methods as in the previous paper, as well as PRS based on 50,633 loci. They again found that PRS outperformed all machine learning methods, with an AUC of 0.90 in the German test data.

Overall, these studies suggest that PRS is generally a more accurate method than machine learning for predicting complex diseases like CAD, although the performance may vary depending on the specific disease and the methods used.

A comparative analysis of these approaches is also presented as a table given below:

Table 5.
A comparative analysis of methods that calculate PRS.

Ref.	Method	Data	Disease	Finding
[39]	DL-based model (DeepPRS) for scoring diseaserisk	Genome-wide genotype data from UK Biobank	Alzheimer, inflammatory bowel, type 2 diabetes, breast cancer	DeepPRS outperformed otherstate-of-the-art methods (prun- ing and thresholding method, and lasso model)
[40]	Compared ML (SVM and RF) and DL methodsto classical PRS calculation	Simulated GWAS data	NA	ML methods achieved more con-sistent results than PRS
[41]	Compares PRSand ML (ANN) methods for cases/controls classification	Simulated geno-typed data	NA	ML (ANN) outperformed thePRS tools in terms of classifica- tion accuracy.
[42]	Compares PRS and ML meth-ods for predict-ing CAD	German case-control data	Coronary artery disease(CAD)	PRS outperformed all ML meth-ods
Proposed PRSNN	DL for calculat-ing PRS	genotype data from OpenSNP	Mental dis-eases suchas bipolar disorder, de-pression, andpanic disorder	PRSNN performed better than traditional method for all three mental diseases

4. Discussion

In this study, we meticulously assessed the PRS for each individual utilizing two distinct methods: the traditional method and a newly proposed approach termed PRSNN. Both methodologies were scrutinized for their ability to portray the distribution of PRS values accurately, with outcomes indicating a normal distribution. This implies that while the majority of individuals exhibit an average risk of developing bipolar disorder, a subset displays either heightened or reduced risks contingent upon their PRS. The traditional method involves assigning weights to individual genetic variants based on their effect sizes as deduced from PGC summary statistics data, then aggregating the weighted scores for all variants. However, this method may not account for the complex interactions and dependencies among the variants and may suffer from overfitting or underfitting issues, and explains only a small fraction of trait variance, andtheir discriminative ability is low in the general population.

The technique PRSNN is a novel approach that determines PRS by employing weight vectors that estimate the relevance of each SNP, which were derived using a deep learning model. This method can handle high-dimensional data and capture non-linear relationships among the variants. It could reduce the sample size.

Our study has some limitations that should be addressed in subsequent research. However, a limitation of this study is that we did not validate our results with external data from different ancestries because the currently experimented data, i.e., PGC data and the genotypes from openSNP, are mostly of Europeanorigin, which may limit the applicability and reliability of our findings. In the future, we aim to explore other populations or ethnicities, such as the Arab, and compare the results to those of this study, using machine learning, and we will include the phenotypes of the individuals to understand how the PRS and the genotype influence the disease risk more than the phenotypes or the genotypes alone, this will help to understand how the PRS and the genotype influence the disease risk more than the phenotypes or the genotypes alone.

Responses (such as disease outcomes) are intricately linked to genetic data, which are sophisticated. Research-based case-control studies using PRSs have shown that the scores and disease status are positively correlated. These studies have also seen widespread use of PRSs in research investigations. That being said, there is still much to learn about how the genotype-environment interactionaffects the functionality and application of PRS models.

In conclusion, the PRSNN method holds promise for improving disease predic- tion and understanding the genetic underpinnings of complex diseases. However, further validation and exploration of its potential across diverse populations is needed. The use of machine learning techniques will be instrumental in this endeavour, potentially leading to significant advancements in genetic studies.

5. Conclusion

The PRSNN method, as a novel proposed approach for calculating PRS, has shown promising results in disease prediction. It leverages machine learning, which can analyze large sets of genomic data, making it an effective tool for considering interaction and nonlinear effects. However, the study has some limitations, like a lack of validation with external data from different ancestries. Future work will focus on exploring other populations or ethnicities using machine learning. An important area that still needs to be investigated is how genotype- environment interactions affect the functionality and applicability of PRS models. Leveraging machine learning techniques will play a crucial role in this undertaking, potentially resulting in substantial progress within the field of genetic research.

Funding:

The King Salman Center for Disability Research provided funding for this work through Research Group no [KSRG-2023-438].

Acknowledgements:

The authors would like to acknowledge the Artificial Intelligence Centre at AlFaisal University. The authors extend their appreciation to the King Salman Centre for Disability Research for funding this work through Research Group KSRG-2023-438.

Copyright:

References

[1] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219–1224, 2018.

[2] Teresa Morera-Herreras, Cristina Miguelez, Asier Aristieta, Mar´ıa Torrecilla, JA Ruiz-Ortega, and Luisa Ugedo. Cannabinoids and motor control of the basal ganglia: Therapeutic potential in movement disorders. *Cannabinoids in health and disease*, pages 59–92, 2016.

[3] GBD 2019 Mental Disorders Collaborators et al. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150, 2022.

[4] Elizabeth Reisinger Walker, Robin E McGee, and Benjamin G Druss. Mortal- ity in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA psychiatry*, 72(4):334–341, 2015.

[5] Martin Knapp and Gloria Wong. Economics and mental health: the current scenario. *World Psychiatry*, 19(1):3–14, 2020.

[6] Zeyu Meng, Huize Chen, and Shengxi Meng. The roles of tetram- ethylpyrazine during neurodegenerative disease. *Neurotoxicity Research*, 39:1665–1677, 2021.

[7] Ronald C Kessler. The epidemiology of dual diagnosis. *Biological psychiatry*, 56(10):730–737, 2004.

[8] Patrick W Corrigan and Amy C Watson. Understanding the impact of stigma on people with mental illness. *World psychiatry*, 1(1):16, 2002.

[9] Marc De Hert, Christoph U Correll, Julio Bobes, Marcelo Cetkovich-Bakmas, DAN Cohen, Itsuo Asai, Johan Detraux, Shiv Gautam, Hans-Jurgen Möller,

[10] David M Ndetei, et al. Physical illness in patients with severe mental disorders. i. prevalence, impact of medications and disparities in health care.*World psychiatry*, 10(1):52, 2011.

[11] Annette Erlangsen, Per Kragh Andersen, Anita Toender, Thomas Munk Laursen, Merete Nordentoft, and Vladimir Canudas-Romo. Cause-specific life-years lost in people with mental disorders: a nationwide, register-basedcohort study. *The Lancet Psychiatry*, 4(12):937–945, 2017.

[12] Laura Ibanez, Fabiana HG Farias, Umber Dube, Kathie A Mihinduku- lasuriya, and Oscar Harari. Polygenic risk scores in neurodegenerative diseases: a review. *Current Genetic Medicine Reports*, 7:22–29, 2019.

[13] Jennifer A Collister, Xiaonan Liu, and Lei Clifton. Calculating polygenic risk scores (prs) in uk biobank: a practical guide for epidemiologists. *Frontiersin genetics*, 13:818574, 2022.

[14] Mitchell J Machiela, Chia-Yen Chen, Constance Chen, Stephen J Chanock, David J Hunter, and Peter Kraft. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic epidemiology*, 35(6):506–514, 2011.

[15] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols*, 15(9):2759–2772, 2020.

[16] Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R Robertson, Jason M Torres, N William Rayner, Anthony J Payne, Valgerdur Steinthors-dottir, Robert A Scott, Niels Grarup, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet- specific epigenome maps. *Nature genetics*, 50(11):1505–1513, 2018.

[17] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jack-son, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variantsclustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[18] Andrew T DeWan. Gene-gene and gene-environment interactions. *Genetic Epidemiology: Methods and Protocols*, pages 89–110, 2018.

[19] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torka- mani, and Amalio Telenti. A primer on deep learning in genomics. *Naturegenetics*, 51(1):12–18, 2019.

[20] Xiaopu Zhou, Yu Chen, Fanny CF Ip, Yuanbing Jiang, Han Cao, Ge Lv, Huan Zhong, Jiahang Chen, Tao Ye, Yuewen Chen, et al. Deep learning-based polygenic risk analysis for alzheimer's disease prediction. *Communi- cations Medicine*, 3(1):49, 2023.

[21] Lei Song, Aiyi Liu, Jianxin Shi, and Molecular Genetics of Schizophrenia Consortium Gejman PV Sanders AR Duan J Cloninger CR Svrakic DM Buccola NG Levinson DF Mowry BJ Freedman R Olincy a Amin F Black DW Silverman JM Byerley WF. Summaryauc: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics*, 35(20):4038–4044, 2019.

[22] Nilanjan Chatterjee, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J Chanock, and Ju-Hyun Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, 45(4):400–405, 2013.

[23] Louis Lello, Timothy G Raben, Soke Yuen Yong, Laurent CAM Tellier, and Stephen DH Hsu. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific reports*, 9(1):15286, 2019.

[24] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[25] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[26] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[27] Daniel Sik Wai Ho, William Schierding, Melissa Wake, Richard Saffery, and Justin O'Sullivan. Machine learning snp based prediction for precision medicine. *Frontiers in genetics*, 10:267, 2019.

[28] Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, 166(5):540–556, 2009.

[29] Patrick F Sullivan, Arpana Agrawal, Cynthia M Bulik, Ole A Andreassen, Anders D Børglum, Gerome Breen, Sven Cichon, Howard J Edenberg, Stephen V Faraone, Joel Gelernter, et al. Psychiatric genomics: an updateand an agenda. *American Journal of Psychiatry*, 175(1):15–27, 2018.

[30] Bastian Greshake, Philipp E Bayer, Helge Rausch, and Julia Reda. Opensnp– a crowdsourced web resource for personal genomics. *PloS one*, 9(3): e89204,2014.

[31] Rajagopal Venkatesaramani, Bradley A Malin, and Yevgeniy Vorobeychik. Re-identification of individuals in genomic datasets using public face images.*Science advances*, 7(47): eabg3296, 2021.

[32] Niamh Mullins, Andreas J Forstner, Kevin S O'Connell, Brandon Coombes, Jonathan RI Coleman, Zhen

Qiao, Thomas D Als, Tim B Bigdeli, Sigrid Børte, Julien Bryois, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature genetics*, 53(6):817–829, 2021.

[33] Naomi R Wray, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air, Till MF Andlauer, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, 50(5):668–681, 2018.

[34] Takeshi Otowa, Karin Hek, Minyoung Lee, Enda M Byrne, Saira S Mirza, Michel G Nivard, Timothy Bigdeli, Steven H Aggen, Daniel Adkins, Aaron Wolen, et al. Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular psychiatry*, 21(10):1391–1399, 2016.

[35] Giovanni Di Leo and Francesco Sardanelli. Statistical significance: p value,

[36] 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European radiology experimental*, 4:1–8, 2020.

[37] Gavin Hackeling. *Mastering Machine Learning with scikit-learn.* Packt Publishing Ltd, 2017.

[38] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[39] Wonil Chung. Statistical models and computational tools for predictingcomplex traits and diseases. *Genomics & Informatics*, 19(4), 2021.

[40] Ragip Onur Öztornaci, Erdal Coşgun, Cemil Çolak, and Bahar Taşdelen. Prediction of polygenic risk score by machine learning and deep learning methods in genome-wide association studies. *bioRxiv*, pages 2022–12, 2023.

[41] Jiajie Peng, Jingyi Li, Ruijiang Han, Yuxian Wang, Lu Han, Jinghao Peng, Tao Wang, Jiaoye Hao, Xuequn Shang, and Zhongyu Wei. A deep learning- based genome-wide polygenic risk score for common diseases identifies individuals with risk. 11 2021.

[42] R. Onur Öztornaci, Erdal Coşgun, Cemil Çolak, and Bahar Taşdelen. Predic- tion of polygenic risk score by machine learning and deep learning methods in genome-wide association studies. *bioRxiv*, 2023.

[43] Muhammad Muneeb, Samuel Feng, and Andreas Henschel. An empiri- cal comparison between polygenic risk scores and machine learning for case/control classification. 01 2022.

[44] Damian Gola, Jeannette Erdmann, Bertram Müller-Myhsok, Heribert Schun- kert, and Inke König. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*,44, 01 2020.

DOI: 10.55214/25768484.v8i6.2338