PREDICTING AUTISM SPECTRUM DISORDER FROM GWAS DATA USING FMNN

Quang-Huy Hoang[1], Thi-Duong Vu[1], Trinh-Hoang Vu[2,*]

DOI: http://doi.org/ 10.57001/huih5804.2025.355

ABSTRACT

With the advancement of biotechnology, genetic data has become a valuable resource for assessing disease risks. GWAS aim to identify SNPs associated with complex diseases. However, the predictive power of GWAS remains limited due to the complexity of genetic architectures. This study proposes a machine learning-based approach to improve disease risk prediction, particularly for autism spectrum disorder (ASD). By applying feature selection using XGBoost and employing the FMNN model, the study enhances the effectiveness of PRS prediction. Experimental results on the AGRE autism dataset show that FMMNN outperforms traditional models, achieving over 75% in F1-score and accuracy. The findings confirm that combining machine learning with GWAS and PRS can effectively identify individuals at higher genetic risk of ASD.

Keywords: *Autism, GWAS, SNP, polygenic risk score, fuzzy min-max neural network.*

[1]School of Information and Communication Technology, Hanoi university of industry, Vietnam

[2]VNU University of Science, Vietnam National University, Hanoi, Vietnam

[*]Email: hoangvt06012003@fit-haui.edu.vn

Received: 25/7/2025

Revised: 20/9/2025

Accepted: 28/9/2025

1. INTRODUCTION

Nowadays, scientists can utilize DNA data to predict an individual's risk of developing diseases. Except for somatic mutations, DNA remains stable throughout a person's lifetime. Therefore, genetically associated disease risks can be identified as early as birth. This highlights the significance of genetic risk assessment in preventive medicine. For instance, a 2017 study estimated that approximately 72% of women who inherit a BRCA1 mutation and around 69% of those with a BRCA2 mutation are likely to develop cancer before the age of 80 [1]. Identifying individuals who carry deleterious genetic variants enables healthcare providers to offer lifestyle modification recommendations or implement preventive interventions tailored to their risk level.

For monogenic disorders, estimating an individual's disease risk can often involve simply identifying pathogenic variants in a specific gene. Genetic Linkage Analysis (GLA) has long been employed to locate disease-causing genes based on their co-segregation with genetic markers on chromosomes. This method has proven highly effective in pinpointing mutations responsible for certain single-gene disorders, such as Huntington's disease [2, 3] or breast cancer [4]. However, linkage analysis has shown limited efficacy in addressing complex, polygenic, and common diseases.

Genome-Wide Association Studies (GWAS) have been conducted to identify common single-nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) ≥ 1% that are associated with complex traits and diseases [5-7]. The increasing feasibility of GWAS has been largely driven by the advancement of large-scale SNP genotyping technologies at relatively low cost, enabling the analysis of datasets containing hundreds of millions of SNPs [8, 9].

Nevertheless, GWAS remains limited in its ability to accurately predict polygenic diseases. Even when using SNPs with strong associations to disease traits, the predictive performance is often suboptimal [10, 11]. To improve prediction accuracy, recent studies have focused on selecting informative subsets of SNPs that contribute significantly to disease risk. Modern approaches typically incorporate both biological and statistical criteria, such as filtering out SNPs due to linkage disequilibrium or population stratification effects [8, 12-14].

In addition to biologically driven approaches, various feature selection techniques have been explored to identify informative SNP subsets. These include machine learning-based methods [15-17], threshold-based filtering strategies [18], feature elimination during model training [19-21], and approaches that capture nonlinear interactions [22]. Such methods are increasingly integrated into polygenic risk score (PRS) models [23], which estimate disease risk in a target cohort using GWAS summary statistics derived from an independent discovery cohort [24, 25]. With the continuous improvement in predictive model performance and the availability of larger datasets, polygenic risk scores are increasingly contributing to efforts in genetic risk stratification and hold great potential for widespread clinical application.

This work presents the following key contributions: (i) proposing an integrated approach based on FMNN (Fuzzy Min-Max Neural Network) and XGBoost for disease risk prediction; (ii) enhancing the effectiveness of polygenic risk score estimation; and (iii) implementing data preprocessing and quality control using XGBoost. The remainder of this paper is organized as follows. Section 2 presents relevant background knowledge. Section 3 provides a detailed description of the proposed algorithm. Section 4 outlines the experimental setup and results on the Autism GWAS dataset. The final section offers discussion and conclusions.

2. DNA SEQUENCE ASSEMBLY

2.1. Genomic sequencing and Haplotype reconstruction

The process of genome sequencing involves accurately mapping the nucleotide arrangement in a DNA strand. It plays a critical role in uncovering gene structures and functions, as well as identifying genetic variations associated with diseases. As a cornerstone technology in modern molecular biology and genetics, genome sequencing has enabled significant advances in biomedical research and personalized medicine.

Several sequencing methodologies have been developed, including:

- This classical approach to sequencing involves incorporating ddNTPs that halt DNA synthesis, resulting in length-variable fragments used to infer nucleotide order.

- NGS (Next-Generation Sequencing): a high-throughput technology that enables the parallel sequencing of millions of short DNA fragments, significantly accelerating data generation and reducing cost.

- TGS (Third-Generation Sequencing): a more recent approach that allows for real-time sequencing of single DNA molecules without the need for amplification, using technologies such as nanopore-based sequencing or single-molecule real-time (SMRT) sequencing to directly observe DNA synthesis.

The next step following genome sequencing is haplotype phasing, which involves determining the combination of alleles located in close proximity on the same chromosome that are inherited together from a single parent. Identifying haplotypes not only provides a more accurate representation of an individual's genetic architecture, but also plays a crucial role in studies of complex traits and genetic linkage analysis. Haplotype information enhances the power of association studies, improves imputation accuracy, and offers deeper insights into the inheritance patterns of disease-associated variants.

2.2. GWAS

GWAS are a powerful approach used to screen molecular markers across the entire genome in order to identify genetic variants associated with a specific disease or trait. The typical GWAS design involves comparing two groups: a case group (individuals with the disease) and a control group (individuals without the disease). DNA is extracted from blood samples or buccal cells and analyzed using SNP genotyping arrays. These platforms scan the genome to detect single-nucleotide polymorphisms (SNPs) that occur at significantly higher frequencies in the case group. Such SNPs may serve as important genetic markers linked to the disease or trait under investigation..

GWAS data can be categorized into two types of access levels:

- Summary statistics: This dataset includes aggregated metrics such as p-values, effect sizes (odds ratios for binary traits or β coefficients for continuous traits), along with SNP identifiers and their genomic positions.

- Individual-level data: This dataset contains detailed information for each participant, including subject identifiers, pedigree structure, sex, phenotypic traits, allelic information at each SNP locus, and other relevant covariates.

2.3. Polygenic risk score

The Polygenic Risk Score (PRS) is calculated as the weighted sum of risk alleles, where the weights are derived from effect sizes estimated by GWAS. PRS serves as a relative measure that enables clinicians to assess an individual's genetic predisposition to a particular disease. In PLINK, the PRS is typically computed according to Equation (1):

$$\mathrm{PRS_j} = \frac{\sum_{i}^{N} S_i . G_{i,j}}{P . M_j} \tag{1}$$

where: N - the total number of SNPs considered for individual j; S_i - the effect size of SNP i; $G_{i,j}$ - the number of risk alleles of SNP i observed in individual j; M_j - the total number of non-missing SNPs observed in individual j; P - the ploidy of the sample.

A significant proportion of GWAS results to date have been derived from datasets with disproportionate representation across ethnic groups - approximately 78% of participants are of European ancestry, 10% Asian, 2% African, 1% Hispanic, and less than 1% from all other populations combined [26]. Moreover, polygenic risk scores (PRS) have also been shown to capture the contribution of polygenic effects to phenotypes that may not be detectable through GWAS alone [27, 28].

A 2018 study by Amit V. utilized statistical inference on genome-wide data to investigate millions of frequent genetic variants linked to five widespread health conditions: coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer [29]. In 2019, a large-scale study involving 272 authors utilized GWAS data on breast cancer, representing the largest of its kind to date. The study aggregated data from 69 individual studies, comprising 94,075 case samples and 75,017 control samples. As a result, a set of 313 SNPs was identified as optimal for polygenic risk score (PRS) calculation, achieving an area under the curve (AUC) of 63% with a 95% confidence interval.

In addition to genetic factors, the development of a disease may also be influenced by non-genetic factors such as environmental exposures, lifestyle, and other external variables. Therefore, diseases cannot be fully predicted based on genetic information alone. The genetic contribution accounts for only a certain proportion, typically quantified as heritability (h^2), or more specifically, SNP-based heritability (h^2_{SNP}).

Moreover, population stratification, typically considered a confounding factor in GWAS, can also be leveraged to improve the accuracy of PRS estimation. Polygenic risk scores can be applied to estimate an individual's disease susceptibility based on genotyping technologies. However, PRS cannot provide deterministic predictions for complex common diseases, as genetic factors only account for a portion of disease risk, and PRS captures only a subset of the total genetic contribution. Nevertheless, similar to how clinical medicine utilizes a wide range of probabilistic indicators, PRS plays a meaningful role as part of multivariable prediction algorithms.

3. PROPOSED METHOD

Although PRS is derived from large-scale genomic studies, the complexity of polygenic traits combined with environmental influences poses significant challenges for its widespread diagnostic application. To enhance the clinical utility of PRS, this study proposes an alternative approach: employing a Fuzzy Min-Max Neural Network (FMNN) to reduce the dimensionality of the SNP set.

FMNN is an incremental learning neural network model that partitions the data space using fuzzy hyperboxes (fHBs). It inherits the advantages of reinforcement learning methods and is capable of handling large-scale datasets efficiently. By leveraging the flexibility and generalization ability of fuzzy logic, FMNN offers a promising framework for feature selection and classification in high-dimensional genetic data.

3.1. Fuzzy hyperbox-fuzzy membership degree

A fuzzy hyperbox B_j is a region in the n-dimensional sample space, defined by its minimum point V_j and maximum point W_j. The membership degree b_j of a data point to the hyperbox is defined as in Equation (2):

$$\mathrm{B_j} = \left\{\mathrm{A_h}, \mathrm{V_j}, \mathrm{W_j}, \mathrm{b_j^{A_h}}\right\} \tag{2}$$

where: $A_h = (a_{h1}, a_{h2}, \ldots, a_{hn}) \in I^n$, $(h = 1, 2, \ldots, m)$ is the training instance indexed by h; $b_j^{A_h}$ is the membership function, the membership degree b_j of the training sample A_h to the fuzzy hyperbox B_j is computed according to Equation (3):

$$\mathrm{b_j^{A_h}} = \frac{1}{n}\sum_{i=1}^{n}\left[1 - f(a_{hi} - w_{ji}, \gamma) - f(v_{ji} - a_{hi}, \gamma)\right] \tag{3}$$

where: γ is the sensitivity parameter, used to reduce the membership value b_j when the training sample A_h lies outside the fuzzy hyperbox, is computed according to Equation (4):

$$f(x,y)=\begin{cases}1, & xy>1\\ xy, & 0\le xy\le 1\\ 0, & xy<0\end{cases} \tag{4}$$

3.2. Learning algorithm

The learning algorithm consists of expansion and contraction steps to iteratively adjust fuzzy hyperboxes within the sample space.

Let the training set D contain m data samples, with A_h denoting the hth training instance. The learning procedure in FMNN includes the following three main steps:

- Initialization of fuzzy hyperboxes
- Creation and expansion of fuzzy hyperboxes
- Overlap checking and contraction adjustment

Steps 2 and 3 are repeated for each sample in the training set until cluster stability is achieved. Stability is defined as the condition in which all minimum and maximum points of the hyperboxes remain unchanged across two consecutive iterations in the same order.

4. EXPERIMENTS AND EVALUATION

4.1. Experimental dataset

The Autism GWAS SNP dataset [31], related to autism spectrum disorder, contains genome-wide genetic information, including chromosomal positions and minor allele frequencies (MAF). The genotypic dataset consists of 399,147 rows (SNP genotypes) and 8 columns (see Table 1).

In addition, the phenotypic and pedigree data - including family structure and disease status - are provided separately (see Table 2).

Table 1. Structure of the GWAS dataset

Attribute	Description
CHR	Chromosome number indicating where the SNP is located
SNP	Identifier or name of the SNP
BP	Exact base-pair position of the SNP on the chromosome
CM	Genetic distance (in centiMorgans) between markers on the chromosome
A1	Reference (major) allele
A2	Alternative (minor or variant) allele
MAF	Minor Allele Frequency - frequency of the less common allele in the population
NCHROBS	Number of chromosomes observed at the SNP locus

Table 2. Structure of Pedigree and Phenotype data

Attribute	Description
FID	Family ID - a unique identifier for each family
IID	Individual ID - a unique identifier for each individual within the family
FAT	Paternal ID - unique ID of the individual's father
MAT	Maternal ID - unique ID of the individual's mother
Sex	Sex - encoded as 1 for male, 2 for female
Phenotype	Disease status - encoded as 1 for affected, 2 for unaffected

4.2. Data preprocessing with XGBoost

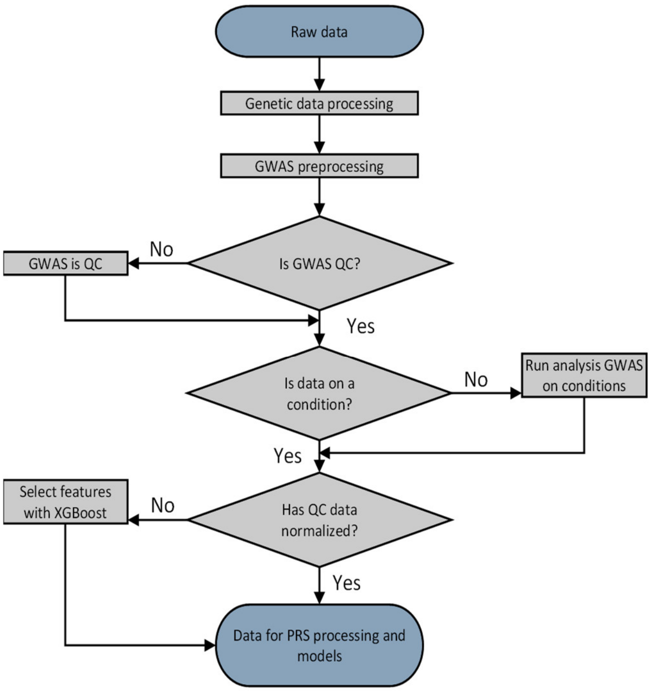


Figure 1. Data quality control pipeline using XGBoost

The predictive performance of Polygenic Risk Scores (PRS) heavily depends on the quality of both the base (discovery) and target datasets. GWAS data must undergo thorough quality control to eliminate technical biases and low-quality variants. In this study, we applied XGBoost as a preprocessing tool to optimize the input feature set (see Figure 1).

Quality control was conducted based on several key criteria:

- Genotyping rate > 0.9, to retain SNPs with sufficient call rates,

- Sample missingness < 0.01, to exclude individuals with excessive missing data,

- Hardy-Weinberg equilibrium (HWE) with $P > 10^{6}$, to ensure population genetic balance,

- Elimination of highly correlated SNPs with pairwise $r^2 > 0.5$, to reduce multicollinearity and redundancy in the feature space.

4.3. Objectives and Evaluation Metrics

The evaluation of model performance is based on the following metrics:

- Accuracy. The proportion of correctly predicted cases among all predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

- Precision. The proportion of true positive predictions among all positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

- Recall. Recall (Sensitivity): The proportion of true positive predictions among all actual positive cases:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

- F1-score. The harmonic mean of precision and recall, providing a balanced measure of both:

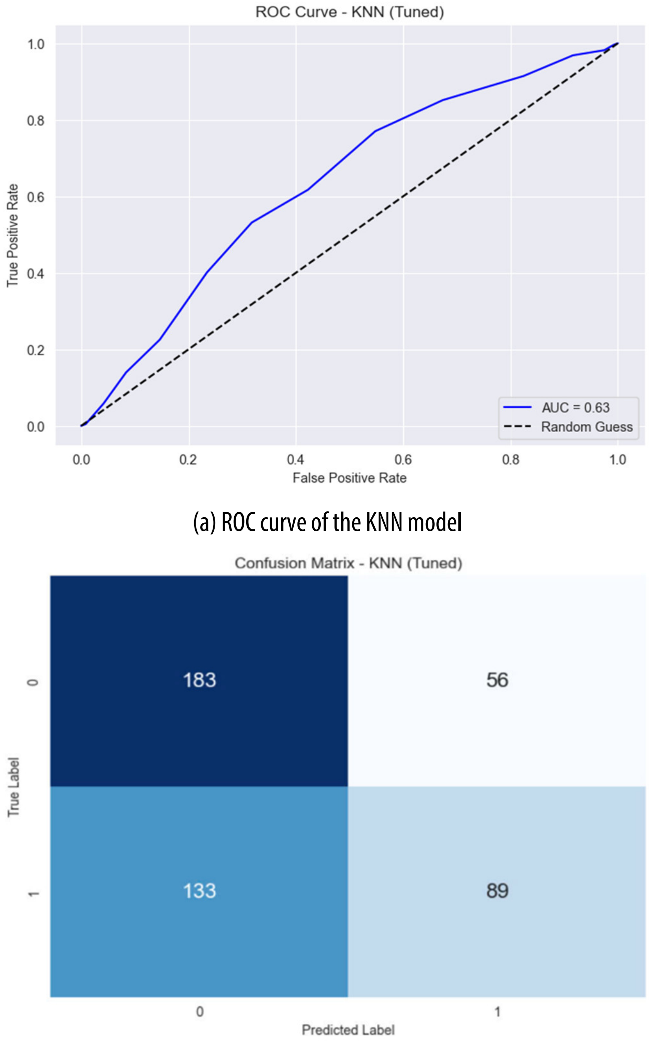
$$\text{F1} = 2.\frac{\text{Precision}\,.\,\text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

In addition to the aforementioned metrics, the study also employed the CM (Confusion Matrix) and the ROC (Receiver Operating Characteristic) curve to evaluate the performance of the binary classification models.

4.4. Experimental results

The experiments were conducted using two computing environments. The first consisted of two personal desktop machines equipped with Intel Core i9-13900K CPUs, NVIDIA GeForce RTX 4090 GPUs, 64GB DDR5 RAM (2 x 32GB), 1TB PCIe Gen 4.0 NVMe SSDs, and ASUS ROG Z790-E GAMING WIFI motherboards. These systems were used to run machine learning models on the raw GWAS dataset over a period of one week.

The second setup involved a laptop configured with a 12th Gen Intel Core i7-12700H CPU, NVIDIA GeForce RTX 3050 GPU, and 32GB RAM. This system was used to perform data quality control, feature selection, and model training, with a total processing time of approximately 15 hours.



(a) ROC curve of the KNN model

(b) Confusion matrix of the KNN model

Figure 2. ROC curve and confusion matrix of the KNN model

Table 4 presents a comparison of model performance before and after data filtering, where 95% of the most informative SNPs were retained. Feature selection was used to eliminate low-importance variants, reduce noise, and focus on the most relevant genetic markers. This led to improved model performance across various evaluation metrics.

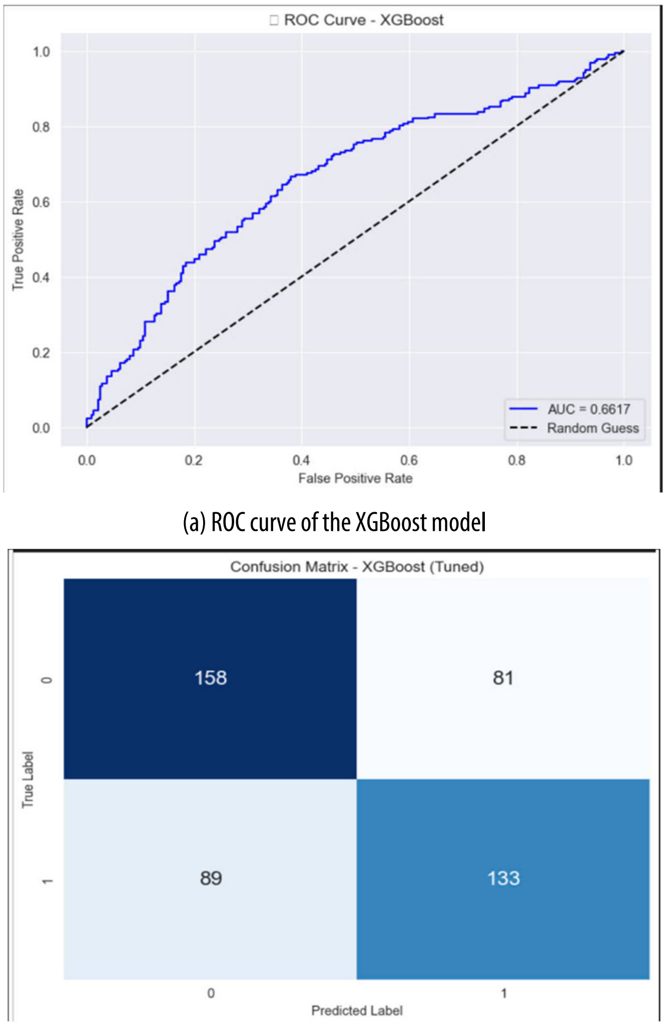
For the KNN algorithm, all performance metrics improved after feature selection, indicating that the filtering process helped the model focus on the most relevant features. This reduced the influence of low-value

Table 3. Comparison of Model Performance: KNN, XGBoost, and FMNN

Model	Initial results				After feature selection (95%)			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
KNN	0.4857	0.4087	0.4454	0.4887	0.6138	0.4009	0.4850	0.5900
XGBoost	0.5221	0.4667	0.4936	0.5073	0.6215	0.5991	0.6101	0.6312
FMNN	0.6335	0.5708	0.5721	0.6304	0.7276	0.7821	0.7539	0.7589

or noisy features and led to an overall enhancement in model accuracy.

XGBoost showed a significant improvement after feature selection. The filtering process helped the model avoid overfitting and enhanced its generalization ability on the test dataset.



(a) ROC curve of the XGBoost model

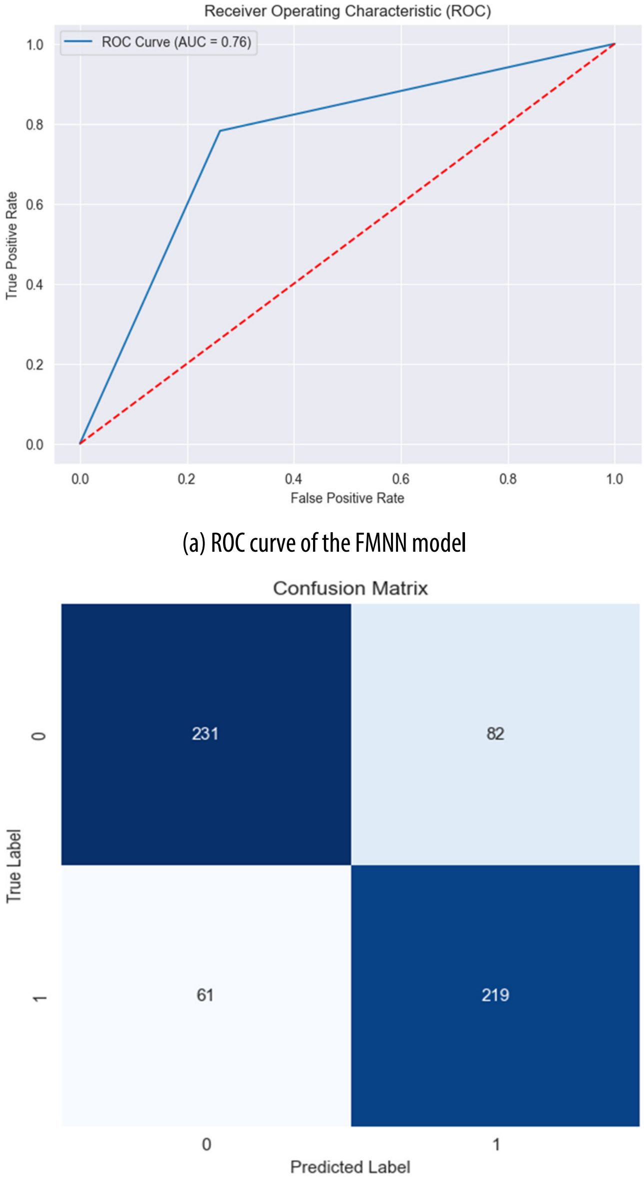
(b) Confusion matrix of the XGBoost model

Figure 3. ROC curve and confusion matrix of the XGBoost model

The FMNN algorithm demonstrated relatively strong performance from the outset. After feature selection, the model maintained its advantage, with a slight increase in evaluation metrics. Since FMNN inherently adjusts its structure to identify the most relevant regions in the data space, the feature selection process did not significantly alter its performance.

The results across all models indicate that the overall prediction rates were not particularly high, which can be attributed to the complex nature of autism spectrum disorder (ASD). Prior studies have shown that environmental factors contribute approximately 9 - 36% to autism risk. These include pre- and perinatal factors such as advanced parental age, birth complications involving hypoxia, preterm birth, and maternal obesity. Additionally, nutritional influences during pregnancy and exposure to environmental toxins have also been implicated [30].

Therefore, with a dataset containing only genetic factors, the models were inherently limited in their ability to fully capture the risk of autism in the studied population.



(a) ROC curve of the FMNN model

(b) Confusion matrix of the FMNN model

Figure 4. ROC curve and confusion matrix of the FMNN model

5. CONCLUSION

This study proposed a machine learning-based approach to improve the prediction accuracy of autism risk by integrating GWAS data with the FMNN model. Experimental results on the AGRE dataset demonstrated that FMNN, when combined with XGBoost for feature selection, outperformed traditional methods in terms of both accuracy and F1-score, achieving scores above 75%. These findings highlight the potential of machine learning in enhancing risk prediction for complex polygenic disorders such as autism.

Despite the promising results, there are several limitations that should be addressed in future work: *(i)* The current dataset focuses solely on genetic factors without considering environmental variables, which play a crucial role in the development of autism; and *(ii)* The sample size and ethnic diversity within the dataset are limited, affecting the generalizability of the model.

Future directions for this research include: *(i)* Expanding the study to incorporate more diverse datasets that combine both genetic and environmental factors; *(ii)* Optimizing the FMNN model to reduce computational time while maintaining high accuracy; and *(iii)* Applying the proposed approach to other complex polygenic diseases, such as diabetes, cardiovascular disorders, and schizophrenia.

REFERENCES

[1]. Kuchenbaecker K., et al., "Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers," *JAMA*, 317, 2402-2416, 2017.

[2]. Wexler Nancy S., et al., "Homozygotes for Huntington's disease," *Nature,* 326.6109, 194-197, 1987.

[3]. Gusella James F., "Location cloning strategy for characterizing genetic defects in Huntington's disease and Alzheimer's disease," *The FASEB Journal,* 3.9: 2036-2041, 1989.

[4]. Ford D., et al., "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families," *The American Journal of Human Genetics*, 62, 676-689, 1998.

[5]. Klein R., et al, "Complement factor H polymorphism in age-related macular degeneration," *Science,* (New York, N.Y.) 308, 385-389, 2005.

[6]. Burton P., et al., "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature,* 447, 661-678, 2007.

[7]. Loos R., "15 years of genome-wide association studies and no signs of slowing down," *Nature Communications,* 11, 5900, 2020.

[8]. Visscher P., et al., "10 Years of GWAS Discovery: Biology, Function, and Translation," *American Journal of Human Genetics*, 101, 5-22, 2017.

[9]. Bycroft C., et al., "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, 562, 203-209, 2018.

[10]. Pepe M., et al., "Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker," *American Journal of Epidemiology*, 159, 882-890, 2004.

[11]. Jakobsdottir Johanna, et al., "Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers," *PLoS genetics*, 5.2: e1000337, 2009.

[12]. Wray Naomi R., Michael E. Goddard, Peter M. Visscher, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome research*, 17.10: 1520-1528, 2007.

[13]. Cecile A., J. W. Janssens, Michael J. Joyner. "Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better?." *Clinical chemistry* 65.5: 609-611, 2019.

[14]. Sha Zhendong, Ting Hu, Yuanzhu Chen, "Feature selection for polygenic risk scores using genetic algorithm and network science," in *2021 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2021.

[15]. Klinger Jörn E., et al., "Interaction-based feature selection algorithm outperforms polygenic risk score in predicting Parkinson's Disease status," *medRxiv*, 2021-07, 2021.

[16]. Kulm Scott, Jason Mezey, Olivier Elemento, "Benchmarking polygenic risk score model assumptions: Towards more accurate risk assessment," *bioRxiv*, 2022-02, 2022.

[17]. Privé Florian, et al., "Making the Most of Clumping and Thresholding for Polygenic Scores," *The American Journal of Human Genetics*, , 105, 6, 1213-1221, 2019.

[18]. Hahn G., et al., "A Fast and Efficient Smoothing Approach to Lasso Regression and an Application in Statistical Genetics: Polygenic Risk Scores for Chronic Obstructive Pulmonary Disease (COPD)," *Statistics and Computing*, 31, 35, 2021.

[19]. Pattee J., W. Pan, "Penalized Regression and Model Selection Methods for Polygenic Scores on Summary Statistics," *PLOS Computational Biology*, 16, e1008271, 2020.

[20]. Dickson S., et al., "GenoRisk: A Polygenic Risk Score for Alzheimer's Disease," *Alzheimer's Dementia: Translational Research & Clinical Interventions*, 7, e12211, 2021.

[21]. Peng J., et al., "A Deep Learning-Based Genome-Wide Polygenic Risk Score for Common Diseases Identifies Individuals with Risk," *medRxiv*, 2021.

[22]. Zhao B., F. Zou, "On Polygenic Risk Scores for Complex Traits Prediction," *Biometrics*, 2021.

[23]. Euesden James, Cathryn M. Lewis, Paul F. O'Reilly, "PRSice: Polygenic Risk Score Software," *Bioinformatics*, 31, 9, 2015, 1466-1468, 2015.

[24]. Uffelmann E., et al., "Genome-Wide Association Studies," *Nature Reviews Methods Primers*, 1, 1-21, 2021.

[25]. Sirugo Giorgio, Scott M. Williams, Sarah A. Tishkoff, "The Missing Diversity in Human Genetic Studies," *Cell*, 177, 1, 2019, 26-31, 2019.

[26]. Purcell S., et al., "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder," *Nature*, 460, 7256, 748-752, 2009.

[27]. Wray Naomi R., et al., "Research Review: Polygenic Methods and Their Application to Psychiatric Traits," *Journal of Child Psychology and Psychiatry*, 55, 10, 2014, 1068-1087, 2014.

[28]. Khera A., et al., "Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations," *Nature Genetics*, 50, 9, 1219-1224, 2018.

[29]. Havdahl Astri, et al., "Genetic Contributions to Autism Spectrum Disorder," *Psychological Medicine*, 51, 13, 2260-2273, 2021.

[30]. Autism GWAS Data, *Figshare*, https://figshare.com/articles/dataset/Autism_GWAS_data/14253230. Accessed 21 June 2025.