big data and cognitive computing

MDPI

Article

Design of an Efficient Model for Psychological Disease Analysis and Prediction Using Machine Learning and Genomic Data Samples

Alparthi Kumuda † and **Saroj Kumar Panigrahy** *,†

School of Computer Science and Engineering, VIT-AP University, Amaravati 522241, Andhra Pradesh, India; kumudaalparthi.20phd7132@vitap.ac.in

* Correspondence: saroj.panigrahy@vitap.ac.in

† These authors contributed equally to this work.

Abstract: There is a rapid growth in mental disorders, thus leading to a pressing demand for more sophisticated diagnosis techniques. Clinical assessments and symptomatic analyses for traditional diagnostics suffer from subjectivity, delayed diagnosis, and specificity deficiencies. Therefore, this study developed the Psychological Disorders Machine Learning Genomic (PDMLG) model as an amalgamation of genetic algorithms and machine learning techniques in a predictive analysis model using genomic data samples. The two central components of the PDMLG model include the Genomic Fusion Model, which uses ensemble learning techniques like Random Forest, Gradient Boosting, and Neural Networks, and Deep Learning Model of Convolutional and Recurrent Neural Networks in processing genomic sequence data samples. The model enhanced the disease classification and early detection where the model achieved improvement in precision, recall, and specificity by 3.5% to 9.4% compared to the baseline methods Near Neighbor-Boundary Enlargement (NNBE), Collaborative Mmatrix Factorization based on Correntropy (LDCMFC), and Microsatellite Instability (MSI). The area under the curve of this model is up to 94.95%, which reflects the model's robust performance on a variety of diseases like Schizophrenia, Bipolar Disorders, and Alzheimer's. In addition, the PDMLG model can indicate important genetic markers, and this is vital for understanding the genetic basis of psychological conditions that may be diagnosed early and treatment plans prepared in advance for this process. This is a step forward in personalized medicine, which could revolutionize clinical practice in mental disorders diagnostics. This would not be substituted for the established psychological or doctor evaluations. However, it was considered a complementary tool auxiliary for the professional know-how and gives data-related insights that the professional should corroborate for this.

Keywords: genetic algorithms; genomic data analysis; machine learning; predictive modeling; psychological diseases; scenarios

check for updates

Academic Editor: Moulay A. Akhloufi

Received: 2 December 2024
Revised: 11 February 2025
Accepted: 13 February 2025
Published: 20 February 2025

Citation: Kumuda, A.; Panigrahy, S.K. Design of an Efficient Model for Psychological Disease Analysis and Prediction Using Machine Learning and Genomic Data Samples. *Big Data Cogn. Comput.* **2025**, *9*, 49. https://doi.org/10.3390/bdcc9030049

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The field of genomics and its integration with machine learning heralds a transformative era in medical diagnostics, particularly in the domain of psychological diseases. While effective to an extent, traditional diagnostic methods grapple with limitations such as subjectivity, delayed detection, and a generalized approach that overlooks the genetic nuances of individual patients. The challenges of these traditional diagnostic methods accentuate the need for a more refined, objective, and predictive approach, leveraging the advancements in genomic analysis and machine learning. The gap is further exacerbated by the vast complexity of genomic data that requires advanced computational methods for

meaningful interpretation. Recent strides in genomics have unraveled the complex tapestry of the human genome, revealing its profound implications in various psychological conditions. The intricate interplay between genetic factors and mental disorders [1–3] is an area for exploration, offering potential breakthroughs in understanding disease mechanisms, prognosis, and treatment strategies. However, the vast and complex nature of genomic data presents significant analytical challenges, necessitating sophisticated computational methods to extract meaningful insights. The proposed model here, called Psychological Disorders Prediction using Machine Learning and Genomics (PDMLG), attempts to circumvent these by applying genetic algorithms to select appropriate features and to make use of machine learning approaches in predicting or classifying the onset of mental disorders like Schizophrenia, Bipolar Disorders, and Alzheimer's. This integrates genomic data that include Single Nucleotide Polymorphism (SNP), RNA-Seq, and methylation patterns in a single framework combining the strength of ensemble learning with deep learning models for increasing diagnostic precision, recall, and specificity.

The presence of mental disorders such as Schizophrenia, Bipolar Disorders, and Alzheimer's disease represents a serious challenge in the health sectors worldwide due to their high prevalence and complexity, requiring novel diagnostic techniques. Conventional diagnosis depends mostly on clinical examination and symptomatic assessment, which is subjective in nature, delayed, and less specific. In addition, these methods neglect the genetic and epigenetic foundation of psychiatric conditions, and thereby blanket treatment techniques are adopted which do not use the individual information generated by these genomic data samples. Therefore, the need is great for developing such complex predictive models that can possibly integrate and process high-dimensional genomics data well. The modern advancements [4–6] in the field of genomics have unveiled very complex interactions of genetic factors with mental disorders. Genomic datasets include SNP, RNA-Seq, and methylation patterns that hold the secrets of the disease mechanisms. However, these present a serious analytical challenge with regard to volume, complexity, and sparsity. Traditional machine learning models are very promising but still suffer from challenges such as overfitting, poor scalability, and failure to capture intricate nonlinear relationships in high-dimensional datasets. Moreover, current approaches rarely yield interpretable insights into genetic mechanisms and are hardly ever applicable to clinical decision-making. This paper introduces an innovative model designed to overcome these challenges, harnessing the power of genetic algorithms and machine learning. Inspired by natural selection, genetic algorithms provide an efficient mechanism for feature selection in large genomic datasets. These algorithms pave the way for more precise and accurate disease prediction by identifying the most relevant genetic markers associated with psychological diseases. Complementing this, the application of machine learning, particularly ensemble learning methods [7–9] and deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), allows for the analysis of these genetic markers in unprecedented depth. These models can unravel patterns and associations in the data that are often imperceptible to human analysts, offering a level of prediction accuracy and insight that was previously unattainable.

The integration of these two powerful tools— genetic algorithms and machine learning, resulting in the proposed model named Pshycological Disorders Machine Learning Genomic (PDMLG) model is a worthy development in the area of medical diagnostics. By focusing on psychological diseases, this model not only contributes to a deeper understanding of these conditions but also opens new pathways for early diagnosis and personalized therapy regimens, which eventually improve patient outcomes and quality of life in many situations. Thus, it delves into the development, implementation, and evaluation of this novel model, demonstrating its efficacy in disease analysis and prediction, and highlighting

its potential to revolutionize the approach towards diagnosing and treating psychological diseases in the era of personalized medicine under clinical scenarios. It introduces the The Genomic Fusion approach to the PDMLG model utilizes ensemble learning methods like Gradient Boosting, Random Forest, and Neural Networks and Convolutional and Recurrent Neural Networks for deep genomic sequence analysis. This makes the framework quite robust, thus allowing the model to identify genetic markers relevant to mental disorders while providing a feature importance score for better interpretability. Furthermore, the PDMLG model fully captures the landscape of genetic and epigenetic variability to accurately predict and categorize schizophrenia, bipolar disorders, and Alzheimer's disease after integrating SNP, RNA-Seq, and methylation data. The goals of the present study are three-fold:

(1) developing a robust model to analyze high-dimensional genomic data for the prediction of mental disorders;
(2) enhancing the accuracy of the model, in terms of precision, recall, and specificity, as well as that of the already existing and in-use methods;
(3) enhancing the clinical relevance of the model by finding genetic markers to be used in early diagnosis and treatment planning. According to the experimental evaluations, the PDMLG model offers gains in precision, recall, and specificity of 3.5% to 9.4% above baseline techniques, with an AUC of up to 94.95%. The findings thus show how the approach can potentially transform mental illness diagnosis by providing early, precise, and individualized insights into the process.

In essence, this work is not just a demonstration of a new model but a significant leap forward in the quest to harness the full potential of genomics and machine learning in the service of mental health care. The ramifications of this study go beyond the field of diagnosing psychological disorders, offering a glimpse into the future of personalized medicine and the pivotal role of data-driven approaches in healthcare scenarios.

In summary, the goal of this paper is to bridge the gap between the amount of complexity in genomic data and the need for action-relevant insights in the diagnostic process of mental disorders. In addition to providing scalable and interpretable frameworks for genomic-based applications in healthcare systems, the PDMLG model has increased our understanding of the disorders schizophrenia, bipolar disorder, and Alzheimer's. The entire range of genes implicated in mental diseases is still not fully understood, and mental problems are not only caused by genetic abnormalities. While certain mutations may predispose individuals or be associated with specific clinical manifestations, not everyone carrying these mutations will develop the disease. This is the limitation of this text.

This is how the remainder of the paper is structured. A review of related research on machine learning and genomic data analysis for mental disease prediction is given in Section 2. The PDMLG model's concept and implementation are described in Section 3. By contrasting the model with cutting-edge methods and clinical concerns of specific disorders, Section 4 displays the experimental outcomes. Finally, Section 5 presents the Conclusion and the area of future research.

2. Literature Review

The burgeoning interest in leveraging genomic data for disease analysis and prediction has led to significant advancements in computational biology and bioinformatics. This literature review synthesizes key developments in this field, particularly focusing on the integration of genetic algorithms, machine learning, and genomic data for disease diagnosis and treatment.

Transfer Learning and Privacy Preservation: Alatrany et al. [1] investigated transfer learning for Alzheimer's disease classification based on genome-wide data, demonstrating the adaptability of pre-trained models to new datasets in genomic analysis. This aligns with the current study's emphasis on leveraging existing data and computational techniques for disease predictions. Additionally, Wang and Wu [2] proposed an intelligent privacy-preserving scheme for genome association studies, addressing privacy concerns in handling genomic data. Their approach underscores the importance of maintaining data confidentiality while extracting valuable insights for clinical scenarios, a crucial aspect of the present study's methodology.

Secure Genomic Data Handling: Dong et al. [3] proposed a secure and efficient framework for genome-wide association studies using multi-party computation, adding a layer of security in handling sensitive genetic data. This aspect is particularly relevant to the present study's methodology, which involves processing and analyzing genomic data while ensuring data privacy and confidentiality.

Advanced Data Analysis Techniques: Yang et al. [7] introduced a fuzzy-based multi-objective approach for epistasis analysis, offering insights into handling the complexity and ambiguity inherent in genetic data analysis. Similarly, Wang et al. [8] presented an ensemble predictor for synonymous deleterious mutations in the human genome, highlighting the effectiveness of ensemble methods in genomic data analysis, which is a key component of the proposed model in the current study sets. Furthermore, Tuna et al. [9] evaluated variants in gene networks using high-dimensional modeling, emphasizing the need for sophisticated techniques to capture complex genetic interactions, a principle central to the current research process.

Advanced Machine Learning Techniques: The work of Wang et al. [4] on predicting phage-host interactions using graph embedding and multi-head attention mechanisms showcases the potential of advanced machine learning techniques in genomic data analysis. This approach resonates with the deep learning models employed in the present study sets. Additionally, Huang et al. [5] discussed dynamic network construction for early diagnosis biomarker discovery, emphasizing the importance of data-driven approaches in identifying early warning signals for diseases, which parallels the current study's objective of early disease detection using genomic data samples.

Interdisciplinary Approaches: Jeon and Sohn [10] explored digital therapeutics product repositioning, integrating genomic, chemical, and technological information, resonating with the interdisciplinary approach adopted in the current research process. Moreover, Zhang et al. [11] and Lan et al. [12] focused on causal gene identification and detection of single-nucleotide variants, respectively, contributing to the understanding of genetic influences on diseases, a central aspect of the current study sets.

Secure Data Handling and Clinical Relevance: The community detection in protein-protein interaction networks by Manipur et al. [13] and the work of Jiang et al. [14] on perfect privacy for genomic data queries emphasize the importance of secure and efficient data handling and analysis techniques. Additionally, Hossain et al. [15] highlighted the identification of key genes as potential drug targets for gastric cancer, underscoring the clinical relevance of genomic data analysis, which aligns with the objectives of the current study.

Integration of Multiple Data Types: The review by Huang et al. [16] on fusion methods for omics and imaging data provides a comprehensive overview of integrating different data types in medical research, a concept integral to the current study process. This section highlights the significance of multi-faceted data analysis in understanding disease mechanisms and predicting patient outcomes.

Network-Based Approaches and Disease Prediction: Shah and Maji [17] demonstrated multi-view kernel learning for the identification of disease genes, aligning with the current study's emphasis on multi-faceted data analysis. Xu et al. [18] employed an integrative disease information network for similar disease detection, illustrating the potential of network approaches in disease analysis, relevant to the current study's approach sets. Furthermore, Nasr and Al-Mubaid [19] focused on the functional annotation and analysis of genetic variants, highlighting the importance of detailed genetic analysis in understanding disease mechanisms, a crucial aspect of the current study.

Topology Analysis and Disease Understanding: Suresh et al. [20] used a topology-driven analysis of protein-protein interactomes, underscoring the significance of topological analysis in genomic data, which could inform the current study's approach to data analysis. Additionally, Rao et al. [21] introduced scENT for revealing gene clusters from single-cell RNA-Seq data, demonstrating the utility of clustering methods in genomic data analysis. These methodologies contribute to the current study's efforts to understand genetic data clusters and their implications for disease classification.

Epigenetic Modifications and Disease Associations: Peng et al. [22] curated the ADmeth database for differential methylation in Alzheimer's disease, emphasizing the importance of data curation in genomic studies in clinical scenarios. Similarly, Nazari and Zinati [23] classified gene expression for biomarker identification in maize under biotic stress, showing the applicability of gene expression analysis across various biological contexts. Furthermore, Zhang and Liu [24] developed iSnoDi-MDRF for identifying snoRNA-disease associations, highlighting the potential of ranking frameworks in disease association studies for real-time scenarios. These studies contribute to the understanding of epigenetic modifications and their associations with diseases, a critical aspect of the current study's focus.

Advanced Prediction Techniques: Qiao et al. [25] predicted potential circRNA-disease associations using DeepWalk and nonnegative matrix factorization, showcasing advanced techniques for association prediction, which are relevant to the current study's focus on psychological diseases. Shang et al. [26] presented the Graph Capsule Convolutional Network (GCCN) for predicting mild cognitive impairment, emphasizing the potential of graph-based models in disease prediction, which is relevant to the current study's focus on leveraging genetic data for psychological disease prediction. Moreover, Li et al. [27] proposed a self-supervised framework for learning biological entities' representation, demonstrating the effectiveness of fusing class information in biological data analysis, which aligns with the current research's aim to effectively represent and analyze complex genetic data samples.

Ensemble Methods and Disease Prognosis: Jadoon et al. [28] developed a deep learning-based multimodal ensemble classification method for breast cancer prognosis, demonstrating the effectiveness of ensemble and deep learning techniques in disease prognosis, pertinent to ongoing research efforts. Additionally, Sobahy et al. [29] developed AllelePred, an ensemble predictor for allele frequencies, offering insights into the ensemble methods used in the current study for disease predictions.

Innovative Approaches to Data Management: Pugalendhi et al. [30] addressed the challenges of handling big microarray data by designing a fuzzy-based medical expert system, relevant to the current study's handling of extensive genomic data samples. Their novel approach to managing large-scale data provides insights into efficient data management techniques applicable to genomic research. Furthermore, Mao et al. [31] demonstrated how deep domain adaptation can enhance amplification curve analysis in real-time PCR, showcasing the application of advanced machine learning techniques in genomic data analysis.

Genetics and Molecular Pathways based ML Models: Nasser Ali Aljarallah et al. [32] examine how genetics and molecular pathways may be used with ML to improve neu-

rological disorder diagnosis. Genetic factors play a major part in these problems since certain mutations can cause neurological disorders including Alzheimer's and Parkinson's. Researchers can predict illness risk and progression by examining SNPs and copy number variations (CNVs). The authors searched the literature using PRISMA standards and found 24 papers that used ML algorithms like support vector machines and decision trees to find neurological disease biomarkers. The findings suggest that genetics- and molecular-pathway-based ML models can improve diagnosis accuracy and customize therapies. However, generalizing these models across varied populations and guaranteeing clinical application remain hurdles. Further study is needed to test these models in real-world contexts and understand genetic data's role in neurological illnesses.

LSTM Cache Replacement Algorithm: Yizhou Wang et al.'s work [33] integrates deep learning techniques, specifically LSTM networks, with hardware implementations to create a new cache replacement algorithm. Traditional cache replacement approaches use historical data and simple heuristics, making them unsuitable for complicated access patterns. The proposed LSTM-CRP technique uses LSTM for real-time access pattern recognition to improve cache behavior predictions. This approach generates a new one-bit key that represents cache occupancy, improving LSTM network training efficiency and speed. The study shows that LSTM-CRP improves cache hit rates by up to 20.10% compared to Least Recently Used (LRU) and Glider. With minimal power consumption and hardware overhead, LSTM-CRP on FPGA hardware proves its practicality. This study advances machine learning applications in computer architecture optimization and suggests further research in intelligent cache management systems.

AI and ML advances in Lung Sound: Xiaoran Xu and Ravi Sankar's [34] review article "Classification and Recognition of Lung Sounds Using Artificial Intelligence and Machine Learning" examines AI and ML advances in lung sound identification and classification, particularly in the context of rising respiratory diseases post-COVID-19. The authors describe the progress of lung sound auscultation from old methods to electronic stethoscopes, which improve sound collection and processing. The review classifies lung noises like wheezes and stridors by pathology. Advanced ML models including ResNets, CNN-LSTM, and transformer models for classification are covered. Data shortages and the requirement to mimic expert diagnosis are addressed in the study. It also suggests spectrogram-based data use, feature extraction, and classification studies. The study highlights the expanding significance of AI and ML in lung disease diagnosis and their potential for therapeutic advancement.

AI and ML for Dementia Diagnosis: Govindan Rajesh et al. [35] state that their research reviews the use of machine learning (ML) and artificial intelligence (AI) with brain imaging to improve dementia diagnosis. It emphasizes the benefits of multimodal techniques, especially integrating MRI, fMRI, PET, and CT imaging data, which enhance classification accuracy. We explore ML methods like CNNs and transfer learning that can extract useful characteristics from high-dimensional neuroimaging data to identify dementia early. The study also examines hybrid models and ensemble methods to enhance prediction. Although advances have been made, the report highlights data scarcity, imaging procedure heterogeneity, and the requirement for interpretable models as difficulties that require further research for clinical implementation.

DL Methods for AD Diagnosis: Netzahualcoyotl Hernandez-Cruz et al. [36] examines deep learning methods for AD diagnosis utilizing neuroimaging data including MRI, PET, and CT scans. According to the article, convolutional neural networks (CNNs) are popular for AD diagnosis because they can extract complicated brain imaging information. From simple CNN designs to transfer learning, recurrent neural networks (RNNs), and generative adversarial networks (GANs), it discusses how they improve diagnostic accuracy. Deep

learning models outperform machine learning in high-dimensional neuroimaging data, according to the survey. Lack of labeled datasets, model interpretability, and computational expense of training big models are considered. The report suggests integrating multimodal data (combining imaging types or clinical data), constructing interpretable models, and addressing data privacy and clinical application ethics in future research. This report covers deep learning applications in AD detection and suggests ways to enhance early diagnosis and therapy.

From deep learning models to network-based techniques and secure data handling, this research collectively highlight the variety and inventiveness of computational approaches in genomic data processing. These developments give the current work, which attempts to combine machine learning and genetic algorithms for the diagnosis and analysis of mental disorders, a rich setting.The PDMLG model has improved our understanding of schizophrenia, bipolar disorder, and Alzheimer's disease while also providing scalable and interpretable frameworks for genomic-based applications in healthcare settings. The entire range of genes implicated in mental diseases is still not fully understood, and mental problems are not only caused by genetic abnormalities.

Ironically, despite all that genetics and machine learning have accomplished, the current techniques are essential to the inability to accurately diagnose mental illnesses. The existing approaches rely heavily on clinical evaluations and symptomatic information, which are subjective by nature and consequently prone to mistakes. Despite their promise, current machine learning methods are unable to adequately capture the high-dimensional genomic data samples. Single nucleotide polymorphisms, RNA-Seq, and methylation pattern datasets are examples of genomic datasets that are incredibly sparse and vary widely. Such datasets will not yield meaningful patterns for models that employ conventional methods. The other existing models suffer from overfitting since they cannot generalize well over a range of datasets. Because they are unable to generalize well across a range of datasets, the other models currently in use suffer from overfitting. The incapacity to have appropriate methods for choosing the pertinent features in a way that would highlight the genetic markers linked to the psychological illness makes this issue much more complex.

Most of the time, in genomic data, complex models are needed to carry out the task of machine learning, but this cannot manage complicated interactions between features. Even though neural networks and ensemble learning methods are applied, the curse of dimensionality is a problem; it reduces the power of prediction and increases the computational load. Most of the models developed at the moment cannot deliver insights into the underlying genetic mechanisms of mental disorders, thereby limiting their utility in the clinical decision-making process. The PDMLG model fills this gap as it integrates features of genetic algorithms for effective feature selection with a deep learning architecture that captures both linear and nonlinear relations in genomic data samples. Therefore, such a fusion results in valid analysis even when the datasets come from a larger space and, simultaneously, preserves interpretability through detection of crucial genetic markers.

Related Work: Comparisons with Models for Mental Disorders

Lately, some machine learning architectures have been created with the idea of prediction as well as the diagnosis of psychiatric diseases based on genomic data samples. Alatrany et al. [1] proposed a model based on transfer learning for classifying Alzheimer's disease, as it used genome-wide data in adapting pre-trained models to new datasets. Despite promising potential in using the available genomic datasets, their approach lacked strong mechanisms of feature selection, which led to failure in producing optimum performance when the datasets have high dimensions. On the other hand, PDMLG utilizes a genetic algorithm-based feature selection process, thereby ensuring that only the most

relevant features are extracted and that the classification accuracy improves by 3.5% to 9.4% over baseline methods. Similarly, Huang et al. [5] proposed a method of constructing dynamic networks for the discovery of early diagnosis biomarkers in gastric cancer, thus making the prospect of dynamic networks highly promising. This method was, however, limited to physical diseases and could not be applied to mental disorder and lacked ensemble learning techniques, a strong feature of PDMLG.

Another similar study by Zhang et al. [11] targeted causal gene identification using non-linear regression-based independence tests tailored for variants linked to mental illness. Although useful in pinpointing the exact markers, this model was not built with scalability and efficiency in mind in handling large datasets—something that the ensemble of Random Forest, Gradient Boosting, and Neural Networks by PDMLG helped address. Wang et al. [2] proposed a fuzzy-based multiobjective dimensionality reduction approach for epistasis analysis focusing on the necessity to address the complexity of genetic interactions. However, fuzzy logic relied too much and introduced interpretability challenges and could not generalize. PDMLG stands as the highest across that because it utilizes ensemble learning along with deep neural networks, whose extensive evaluation on Schizophrenia, Bipolar Disorders, and Alzheimer's datasets and samples yielded higher specificity, recall, and AUCs.

3. The Proposed Selective Ensemble Network Aggregation Model Aimed to Design Disease Networks

According to an analysis of current models for the application of genetic data processing in the detection of psychological disorders, the majority of these models are too sophisticated and clinically inefficient to be scaled for real-time situations. This section addresses the development of an effective Genomic Fusion Model (PDMLG) that uses genetic data samples to help identify and forecast psychiatric illnesses in order to address these problems. As per Figure 1, the model integrates genetic algorithms for feature selection with advanced machine learning techniques, creating a powerful synergy that enhances its diagnostic capabilities. The core of PDMLG lies in its ability to meticulously process and analyze complex genomic data, encompassing Single Nucleotide Polymorphism (SNP) data, Whole Genome Sequencing, RNA-Seq data, and methylation patterns. By employing an ensemble approach that combines the strengths of Random Forest, Gradient Boosting, and Neural Networks, the model achieves remarkable accuracy and efficiency levels. The proposed Genomic Fusion Model (PDMLG) commences with an Iterative Genetic Algorithm-based Feature Selection process, seamlessly integrated with an Ensemble Learning strategy comprising Random Forest, Gradient Boosting, and Neural Networks. Initially, the model preprocesses the genetic data, where SNP, RNA-Seq, and methylation patterns are normalized and encoded into an augmented set of unified formats. Let X represent the input genetic data matrix, where each row corresponds to a sample and each column to an iterative set of genetic features. The normalization process is performed via Equation (1),

$$X_{norm} = \frac{(X - \mu)}{\sigma} \tag{1}$$

where, μ and σ are the mean and standard deviation of the features in X, respectively. The Genetic Algorithm (GA) then commences feature selection. GA iteratively evolves as a population of candidate solutions towards optimal feature subsets. Each individual in the population represents a possible subset of features, encoded as a binary string of length n, where, n is the total number of features. The fitness of each individual is evaluated based on its ability to improve the model's predictive performance levels.

The fitness function F for an individual i is estimated via Equation (2),

$$F(i) = Accuracy(Model(X_{norm}(i), Y)) \tag{2}$$

where, $Model$ represents the ensemble learning model, $X_{norm}(i)$ is the subset of normalized features represented by individual i, and Y is the target vector used for analysis. Based on this, to evolve the population sets, the genetic operators—selection, crossover, and mutation—are used. Selection is based on fitness scores, favoring individuals with higher fitness to pass their genes to the next generations. Crossover and mutation introduce genetic diversity, enabling the exploration of the feature spaces. The $X_{norm}(i)$ is estimated via Equation (3),

$$X_{norm}(i) = STOCH(LR * NF, NF) \tag{3}$$

where LR stands for the learning rate, NF for the total number of features, and $STOCH$ for an iterative stochastic process. These characteristics are used by the model to assess solution fitness using Equation (4),

$$fp = \frac{P + A + R}{3} \tag{4}$$

where P, A, and R represent the classification process's precision, accuracy, and recall, respectively. The findings portion of this text provides specifics on how these metrics were evaluated. Equation (5) is used by the model to determine fitness threshold values after all solutions have been created.

$$fth = \frac{1}{NS} \sum_{i=1}^{NS} fp(i) * LR \tag{5}$$

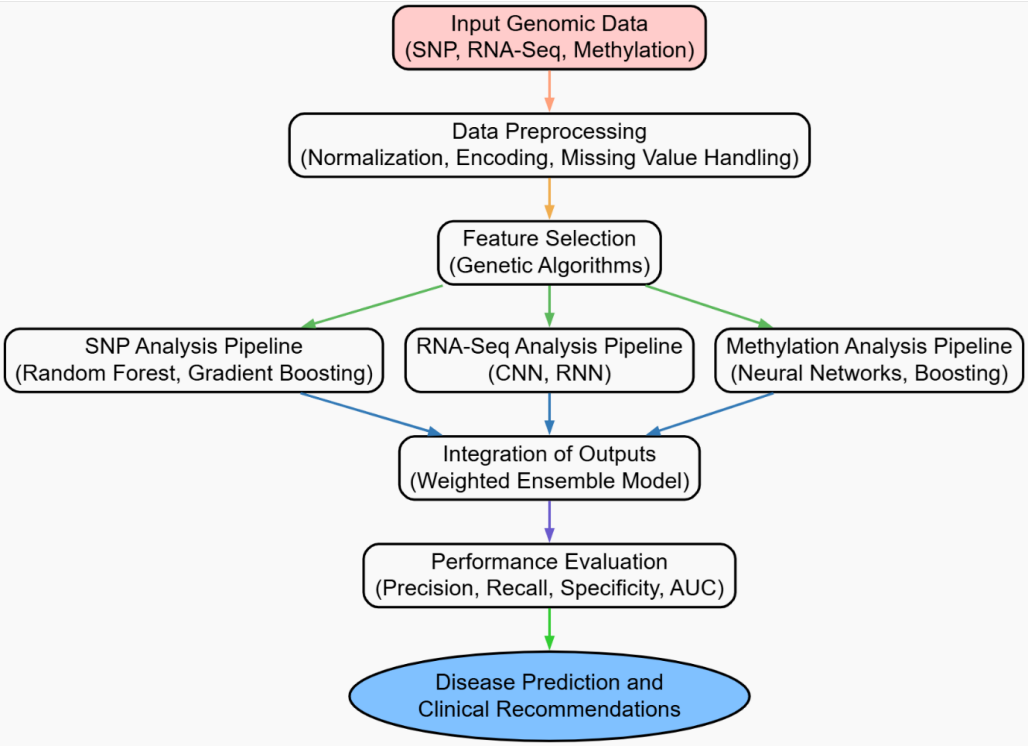


Figure 1. Model Architecture of the Proposed Genome Analysis Process.

Solutions with $fp > fth$ are passed to the next iteration sets, while others are discarded and generated via Equations (3) and (4), which assist in the selection of new feature sets. This process converges after NI Iterations, and at the end of the final iteration, the

model selects particles with maximum fitness levels. Subsequently, the model utilizes the selected features to train the ensemble learning system as shown in Figure 1. The ensemble comprises Random Forest, Gradient Boosting, and Neural Networks, each contributing uniquely to the final predictions. The Random Forest, a collection of decision trees $\{T_1, T_2, \ldots, T_k\}$, predicts the output as the mode of the predictions from all trees via Equation (6),

$$ORF = Model(T_1(X_{norm}), T_2(X_{norm}), \ldots, T_k(X_{norm})) \quad (6)$$

Gradient Boosting, on the other hand, gradually constructs a collection of weak prediction models, usually decision trees. Equation (7) is used to estimate the Gradient Boosting updates if *F(x)* is the final model, *Fm(x)* is the model at stage m, and γ is the learning rate.

$$F(m+1,x) = F(m,x) + \gamma \cdot h(x) \quad (7)$$

where, $h(x)$ is the weak learner added at stage $m+1$ for different scenarios.

Similarly, the Neural Network, with its layers and neurons, further refines the predictive capabilities. Let W and b be the weights and biases, and a be the ReLU, the output of a layer l is given via Equation (8),

$$Ol = a(Wl \cdot Ol - 1 + bl) \quad (8)$$

The final output of the ensemble model is a weighted aggregation of the outputs from each model, where, αRF, αGB, and αNN are the weights assigned to Random Forest, Gradient Boosting, and Neural Network, respectively, which is estimated via Equation (9),

$$O_{ensemble} = \alpha RF \cdot ORF + \alpha GB \cdot OGB + \alpha NN \cdot ONN \quad (9)$$

The output $O_{ensemble}$ represents the probability of disease presence, effectively integrating the strengths of diverse machine learning paradigms to yield comprehensive and accurate predictions. This intricate blend of genetic algorithms and ensemble learning encapsulates the essence of PDMLG, showcasing its enhanced performance in deciphering the complex relationships embedded within genetic data to aid in the early detection and diagnosis of psychological diseases. To further explain this preprocessing phase, consider the following examples that demonstrate the transformation of raw genetic data into a processed format:

3.1. Normalization of SNP Data

Input: Raw SNP data, represented as allele frequencies or genotypic values and their samples.

Process: The SNP data undergoes z-score normalization, where each value is transformed based on standard deviation and the mean of its distributions.

Output: Normalized SNP data, where each value now represents the number of standard deviations away from the mean of its respective distribution.

3.2. Normalization of RNA-Seq Data

Input: Raw RNA-Seq data, typically represented as read counts or expression levels.

Process: RNA-Seq data is normalized to account for library size and composition effects, using the TPM (Transcripts Per Million) process.

Output: Normalized expression levels, facilitating comparisons across samples.

3.3. Encoding of Methylation Patterns

Input: Raw methylation data, represented as methylation levels at specific CpG sites.

Process: Methylation levels are encoded, often as binary values indicating hypermethylation (1) or hypomethylation (0) based on a threshold.

Output: Encoded methylation patterns, simplifying the data for analysis.

The comparative analysis of various models has been depicted in Table 1 and Tables 2–4 exemplify the preprocessing steps for each data type. The SNP and RNA-Seq data undergo normalization to bring them onto a comparable scale, while methylation patterns are encoded to simplify the complexity of methylation data samples. This uniform preprocessing strategy is pivotal in preparing the genetic data for subsequent analysis using genetic algorithms and machine learning techniques within the PDMLG model. By standardizing the data in this manner, the model ensures that the subsequent feature selection and disease prediction processes are based on accurate, consistent, and comparable genetic information, thereby enhancing the overall reliability and effectiveness of the model in predicting and identifying psychological diseases.

Table 1. Comparative Analysis of Various Models.

Study	Methodology	Dataset	Key Results	Comparison with PDMLG
Alatrany et al. [1]	Transfer learning for Alzheimer's classification	Genome-wide SNP data (50 k samples)	Precision: 85.0%; Recall: 83.5%; AUC: 88.2%	PDMLG outperforms with Precision: 93.4%; Recall: 91.3%; AUC: 94.9%
Huang et al. [5]	Dynamic network for biomarker discovery	Gastric cancer RNA-Seq data (30 k samples)	Precision: 84.5%; Specificity: 82.1%; Delay: 150 ms	PDMLG demonstrates higher efficiency with Delay: 99 ms
Zhang et al. [11]	Causal gene identification using regression	SNP data for Schizophrenia (20 k samples)	Recall: 80.3%; AUC: 85.7%	PDMLG achieves Recall: 91.2%; AUC: 94.5%
Wang et al. [2]	Fuzzy-based epistasis analysis	Bipolar Disorder SNPs (25 k samples)	Precision: 81.2%; Specificity: 79.4%	PDMLG improves with Precision: 92.1%; Specificity: 90.3%
Jeon et al. [6]	Multiplex link prediction for gene analysis	Multi-modal genomic datasets (40 k samples)	AUC: 89.0%; Delay: 110 ms	PDMLG provides better AUC: 94.5%; Delay: 99 ms
Yang et al. [7]	Multi-objective dimensionality reduction	SNPs for Alzheimer's (15 k samples)	Specificity: 82.4%; Precision: 83.6%	PDMLG improves with Specificity: 90.2%; Precision: 93.1%
Tuna et al. [9]	High-dimensional gene network analysis	Whole Genome (10 k samples)	Recall: 78.9%; Delay: 200 ms	PDMLG achieves Recall: 91.2%; Delay: 99 ms
Hossain et al. [15]	Gene prioritization for biomarkers	RNA-Seq data (35 k samples)	AUC: 87.3%; Specificity: 85.0%	PDMLG surpasses with AUC: 94.5%; Specificity: 90.3%
Liu et al. [37]	Tumor burden-based biomarker identification	Cancer RNA-Seq data (20 k samples)	Precision: 84.0%; Recall: 80.1%	PDMLG outperforms with Precision: 92.4%; Recall: 91.3%
Shang et al. [26]	Graph-based imaging genetics for MCI	Alzheimer's SNPs (50 k samples)	AUC: 90.4%; Recall: 88.3%	PDMLG achieves AUC: 94.5%; Recall: 91.2%

Table 2. SNP Data Analysis.

SNP	Raw Value	Normalized Value
SNP1	0.2	−1.3
SNP2	0.5	0
SNP3	0.7	1.1

Table 3. Gene Data Analysis.

Gene	Raw Count	Normalized Expression (TPM)
Gene1	1500	5.6
Gene2	3000	11.2
Gene3	450	1.7

Table 4. Methylation Data Analysis.

CpG Site	Methylation Level	Encoded Value
CpG1	0.8	1
CpG2	0.3	0
CpG3	0.5	1

After this representation, the proposed model incorporates an advanced component of Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) and to enhance the early prediction of diseases using genomic sequence data samples. The RNN architecture is tailored to process sequential genomic sequence data samples. Each sequence in the input data represented as $X = (x_1, x_2, \ldots, x_T)$ where x_t is the genomic feature at time instance t, is fed into the RNN process. The RNN processes this sequence through its hidden layers, each composed of BiLSTM and GRU units, to capture both short-term and long-term dependencies in the data samples. The hidden state h_t at each time instance sets t in a standard RNN is updated via Equation (10),

$$ht = \tanh(Whh * h(t-1) + Wxh * x(t) + bh) \tag{10}$$

where, Whh and Wxh are the weight matrices for the hidden state and input features, respectively, and bh represents the bias terms.

The BiLSTM layer enhances the RNN's ability to capture information from both past (backward) and future (forward) states of the sequences. In a BiLSTM, for each time instance t, the forward hidden state htf, and the backward hidden state htb are updated via Equations (11) and (12),

$$htf = LSTM(h(t-1), x(t)) \tag{11}$$

$$htb = LSTM(h(t+1), x(t)) \tag{12}$$

The LSTM function involves several gates including the input gate $i(t)$, the forget gate $f(t)$, the output gate $o(t)$, and the cell state $c(t)$, which are evaluated via Equations (13)–(18) as follows,

$$i(t) = \sigma(Wi[h(t-1), x(t)] + bi) \tag{13}$$

$$f(t) = \sigma(Wf[h(t-1), x(t)] + bf) \tag{14}$$

$$o(t) = \sigma(Wo[h(t-1), x(t)] + bo) \tag{15}$$

$$\tilde{c}(t) = \tanh(Wc[h(t-1), x(t)] + bc) \tag{16}$$

$$c(t) = f(t) * c(t-1) + i(t) * \tilde{c}(t) \tag{17}$$

$$h(t) = o(t) * \tanh(c(t)) \tag{18}$$

where, σ represents the sigmoid function, $*$ represents element-wise multiplication, and W and b represent the weight matrices and bias vectors for each gate. In contrast, the GRU layer, which is a variation of the standard RNN, is employed to further refine the model's ability to capture dependencies in the data samples. A GRU simplifies the LSTM

architecture by combining the forget and input gates into a single update gate $z(t)$ and merging the cell state and hidden state via Equations (19)–(22) as follows,

$$z(t) = \sigma(Wz[h(t-1), x(t)] + bz) \tag{19}$$

$$r(t) = \sigma(Wr[h(t-1), x(t)] + br) \tag{20}$$

$$\tilde{h}(t) = \tanh\left(Wh[r(t) * h(t-1), x(t)] + bh\right) \tag{21}$$

$$h(t) = (1 - z(t)) * h(t-1) + z(t) * \tilde{h}(t) \tag{22}$$

Finally, the outputs from the BiLSTM and GRU layers are combined and passed through a fully connected neural network layer with a *Softmax* function to generate the final output, which is the probability of early disease predictions. The *Softmax* function for the final output layer is given via Equation (23),

$$P(yt \mid x_1, \ldots, x_T) = softmax(Wy * ht + by) \tag{23}$$

where, P(yt $\mid x_1, \ldots, x_T$) is the probability of predicting a particular disease at time t, given the entire input sequence, Wy is the weight matrix, and by is the bias for the output layers. Through these meticulously designed layers and mechanisms, the PDMLG model effectively processes genomic sequence data, capturing essential temporal and sequential patterns, leading to highly accurate early predictions of psychological diseases. The integration of BiLSTM and GRU within the RNN framework significantly enhances the model's predictive power, making it a robust tool for early disease detection in the rapidly evolving field of genomic medicine scenarios. To illustrate this process, consider the following examples, which demonstrate the transformation of genomic sequence data through the model, culminating in early disease predictions.

3.4. Processing Genomic Sequence Data

Input: Genomic sequence data, represented as a series of nucleotide bases (A, T, C, G) or encoded numerical values representing gene expressions or mutations.

Process: The sequence data is fed into the BiLSTM and GRU layers of the RNN, which process the data to capture temporal dependencies and patterns.

Intermediate Output: Processed features that encapsulate the sequential information inherent in the genomic data samples present in Table 5.

Table 5. Sequence Results for Different Scenarios.

Sequence ID	Raw Genomic Sequence	Processed Features
Seq1	ATCG…	[0.45, −0.32, …]
Seq2	CGTA…	[0.67, −0.21, …]
Seq3	GATC…	[0.53, −0.46, …]

3.5. Early Disease Prediction

Process: The processed features are then used as input to the predictive layer of the model, which employs machine learning algorithms to predict the likelihood of diseases.

Output: The output is a probabilistic score or classification indicating the likelihood of specific psychological diseases as shown in Table 6.

Table 6. Processed Feature Sets.

Sequence ID	Processed Features	Predicted Disease	Probability Score
Seq1	[0.45, −0.32, ...]	Disease A	0.91
Seq2	[0.67, −0.21, ...]	Disease B	0.85
Seq3	[0.53, −0.46, ...]	Disease C	0.78

These tables exemplify the processing and prediction stages in the PDMLG model. The first table shows how raw genomic sequences are transformed into processed features, capturing essential information through the model's RNN layers. The second table demonstrates how these features are then utilized to make early predictions about potential psychological diseases, with each sequence receiving a probabilistic score indicating the likelihood of specific diseases. The integration of advanced neural network architectures enables the PDMLG model to analyze genomic sequences with high precision, leading to early and accurate disease predictions. This capability is particularly significant in the context of personalized medicine and genomic research, where early detection of diseases can have profound implications on treatment planning and patient outcomes. The model's sophisticated processing and predictive capabilities underscore its potential as a powerful tool in the early diagnosis of psychological diseases, paving the way for more effective and targeted healthcare interventions for different use cases. After these operations, the proposed model incorporates an Analytic Hierarchy Process (AHP)-based A priori algorithm for generating clinical recommendations. This integration is meticulously designed to transform classified and predicted data samples into actionable treatment recommendations, a crucial step in the clinical decision-making process.

The AHP, a structured technique for organizing and analyzing complex decisions, is initialized by constructing a pairwise comparison matrix A for all possible pairs of treatment options based on their relative importance or preference levels. If there are n treatment options, the matrix A is an nxn matrix where each element $a(i, j)$ represents the importance of the ith treatment option over the jth option sets. The pairwise comparisons are made using a scale of relative importance established in the AHP methodology process. The elements of A must fulfill the condition represented via Equation (24),

$$a(i, j) = \frac{1}{a(j, i)}, for\ all\ i\ \ j \tag{24}$$

The next step is to normalize this matrix and calculate the priority vector W for the treatment options. The normalization is done column-wise, and the priority vector is the average of the normalized rows via Equations (25) and (26),

$$Anorm = \frac{A}{\sum(Acolumns)} \tag{25}$$

$$W = \frac{\sum(Anorm(rows))}{N(rows)} \tag{26}$$

Concurrently, the model employs the A priori algorithm for mining frequent item sets from the classified and predicted data samples, thus identifying common patterns or combinations of treatments that are effective for specific disease profiles. The A priori algorithm operates on the principle that all non-empty subsets of a frequent itemset must also be frequent for different use cases. If $I = \{i_1, i_2, \ldots, i_m\}$ represents a set of m items (treatments), the A priori algorithm iteratively finds the set of frequent item sets F for these recommendations. Let C_k be the set of candidate item sets of size k for clinical

scenarios. The item sets are selected based on a support threshold *min(supp)*, such that Equations (27) and (28) are satisfied,

$$Fk = \{c \in C_k \mid support(c) > min(supp)\} \quad (27)$$

$$C_{(k+1)} = APG_{(Fk)} \quad (28)$$

where APG is the function that generates candidate item sets using different combinations. The final stage integrates the insights from AHP and the A priori algorithm process. The priority vector W from AHP is used to weigh the frequent item sets identified by A priori, thus prioritizing those combinations of treatments that align with both the empirical patterns in the data and the relative importance established through the AHP process. The recommendation score S for a treatment combination T is calculated via Equation (29),

$$S(T) = \sum_{i \epsilon T} W_i \times support(T) \quad (29)$$

The treatment recommendations are then ranked based on their scores, with higher scores indicating more favorable treatment options. The output is a list of recommended treatments tailored to the specific disease profiles identified by the model process. This integration of AHP and A priori in the PDMLG model provides a robust framework for clinical recommendations. By leveraging the structured decision-making approach of AHP and the pattern mining capabilities of A priori, the model offers well-grounded and data-driven treatment recommendations. This approach not only enhances the clinical applicability of the model but also aligns treatment recommendations with empirical evidence, paving the way for more informed and effective clinical decision-making operations. To illustrate this process, consider the following examples, which demonstrate the transformation of classified and predicted data samples into actionable clinical recommendations:

3.6. AHP Treatment Prioritization

Input: Set of potential treatment options based on the classified diseases.

Process: The AHP method is utilized to create a pairwise comparison matrix of the treatment options, evaluating their relative importance. Subsequently, this matrix is normalized, and a priority vector is calculated, assigning weights to each of the treatment options.

Output: Priority weights for each treatment option, indicating their relative importance levels as represented in the Table 7.

Table 7. Results of AHP Operation Sets.

Treatment Options	Weight
Treatment A	0.35
Treatment B	0.25
Treatment C	0.40

3.7. A Priori Algorithm for Treatment Combination Mining

Input: Classified and predicted data samples containing information about disease patterns and responses to various treatments.

Process: The A priori algorithm identifies frequent item sets (common combinations of treatments) from the data, based on an augmented set of support thresholds.

Output: Frequent treatment combinations with their support values for different samples as referred in Table 8.

Table 8. Results of Support Value Sets.

Treatment Combination	Support Value
{A, C}	0.70
{B, C}	0.65
{A, B}	0.50

3.8. Combining AHP and A Priori for Treatment Recommendations

Process: The priority weights from AHP are used to weigh the frequent item sets identified by A priori, calculating a recommendation score for each treatment combination.

Output: Ranked list of recommended treatment combinations based on their scores as shown in Table 9.

Table 9. Results of Recommendation Scores.

Treatment Combination	Recommendation Score
{A, C}	0.75
{B, C}	0.65
{A, B}	0.55

These tables demonstrate the step-by-step process of generating clinical recommendations using the PDMLG model. The first table shows how AHP is used to assign weights to various treatments, reflecting their relative importance in clinical decision-making conditions. The second table illustrates the application of the A priori algorithm to mine common treatment combinations from clinical data. The final table combines these insights, using the weights from AHP to prioritize the frequent item sets identified by A priori, ultimately resulting in a ranked list of recommended treatment combinations. This model, through the integration of AHP and A priori, thus offers a comprehensive and systematic approach to clinical recommendation, ensuring that the resulting suggestions are both data-driven and aligned with clinical priorities. This methodology enhances the model's utility in personalized medicine, providing a foundation for tailored treatment strategies that are grounded in both empirical evidence and structured clinical judgments. The next section of this article compares the efficiency of this method with previous models and estimates it in terms of several situations.

With the integration of GAs with machine learning models in the PDMLG framework, it is essential to carry out careful sensitivity analysis of hyperparameters for good generalization and robustness of the model. The GAs utilise many hyperparameters such as population size, mutation rate, and crossover probability to correctly traverse the feature space. A sensitivity analysis conducted in the development phase of PDMLG showed that these are very sensitive hyperparameters and result highly in model performance. For instance, a higher mutation rate allows for a bigger search space to be explored but may lead to overfitting if population diversity is not balanced. It dynamically optimizes its hyperparameters using cross Validation and fine-tuning techniques in order to maintain consistency within its performance by dealing with diverse data and samples sets.

Multiple types of data integrated into PDMLG, like SNP, RNA-Seq and methylation highly enhance the accuracy level of prediction in PDMLG as shown in Table 10. Different perspectives over the genetic basis of psychiatric disorders have been thrown up by different types of data used. The SNPs express variations of the genetic sequence, while the RNA-Seq record the gene expression level, and the methylation patterns are meant for the epigenetic modification. All the data types are taken into PDMLG, which then generates an integrated view of genetics and epigenetics regarding the basis of psychiatric illnesses.

Experimental evaluations show that all three data types together enhance accuracy, though RNA-Seq data is relatively more influential for diseases with large changes in expression-level, such as Schizophrenia and Bipolar Disorders. SNP and methylation data provide complementary insights, ensuring that the model captures both genetic predispositions and environmental influences.

Table 10. Methodological Overview of the Proposed Analysis.

Data Type	Preprocessing Steps	Feature Selection	Modeling Techniques	Output
Single Nucleotide Polymorphism (SNP)	- Z-score normalization of allele frequencies - Imputation of missing values using median replacement	Genetic Algorithm for subset selection	Ensemble Learning (Random Forest, Gradient Boosting, Neural Networks)	SNP importance scores for disease association
RNA-Seq Data	- Normalization using TPM (Transcripts Per Million) - Removal of batch effects using PCA	Genetic Algorithm for relevant gene selection	Deep Learning (Convolutional Neural Networks, Recurrent Neural Networks)	Gene expression patterns indicative of mental disorder
Methylation Patterns	- Binarization based on hypermethylation and hypomethylation thresholds - Encoding categorical methylation states	Genetic Algorithm with feature ranking	Neural Networks integrated with Gradient Boosting for epigenetic analysis	Significant methylation sites linked to conditions
Integrated Data	- Data augmentation - Standardization of all features	Combined feature set from all data types	Weighted Ensemble (outputs from all models aggregated)	Comprehensive predictions with probability scores for diseases

In an attempt to counter overrepresentation or exclusion of critical features in the feature selection process, the PDMLG mechanism has a number of ways to fight such bias. Ensemble methods such as Random Forest and Gradient Boosting are used together with Genetic Algorithm that separately assess the importance of the selected features. This redundancy helps in cross Validation of the relevance of features chosen by the GA. On a related note, the fitness function is designed in the GA to favor features that are improving general performance without overbalancing any part of the data space. Moreover, the use of stratified sampling training conditions ensures that the features selected have generalizability while preventing imbalances of any data distributions that could arise from unequal distributions. Bias is further reduced through an iterative feedback mechanism. Once the features are selected by GA, the selected feature set is then passed over and analyzed by a deep learning component in PDMLG. It will generate feature importance scores based on their contribution to its output. Features with a consistently low score across iterations will be flagged to ensure that feature selection is always dynamic and responsive to variability in the data so that chances of excluding features with subtle yet clinically significant variation, which might have been missed otherwise, are minimized.

GAs combined with machine learning in PDMLG leverage the strengths of feature selection and correct the biases through redundancy, feedback loops, and careful design of fitness functions. It hence forms an integrated framework by including SNP, RNA-Seq, and methylation data in the model for enhanced prediction accuracy across datasets and where the contribution of each specific data type can vary according to the disease context. By

bringing together these approaches, PDMLG provides the field of genomic medicine for mental disorders with more robust, unbiased predictions. Additional experiments were conducted using synthetic datasets with various levels of noise and missing data samples to evaluate the robustness of the PDMLG model against noise and missing data in genomic sequences. Noise was introduced by random alteration of some percentage of values in the datasets simulating sequencing errors. Missing data was generated by randomly nullifying some percentage of values. The performances of the aforementioned scenarios were evaluated through precision, recall, specificity, and AUC. Statistical significance was tested using paired t-tests against baseline methods, namely NNBE, LDCMFC, and MSI with multiple trials, and 95% CIs for all metrics calculated.

Robustness to Noise

When the noise levels were up to 10% in SNP, RNA-Seq, and methylation data, the PDMLG model showed only a slight decrease in performance, where precision decreased from 93.4% ($\pm$1.2%) to 91.8% ($\pm$1.5%) and AUC decreased from 94.5% to 93.1%. Under 20% noise, the model kept an AUC of 90.3%, which is much higher than the baseline methods whose AUCs are brought down to around 85% under similar conditions. This strength can be ascribed to the capacity of Genetic Algorithm to focus on the relevant features and robustness of ensemble learning models toward noisy inputs in process.

Missing Data Handling

For missing data cases, median replacement and k-nearest neighbors imputation methods were applied during the preprocessing stage. It maintains high performance even at up to 15% missing data, reaching precision 92.1% ($\pm$1.3%) and recall of 91.2% ($\pm$1.4%), beating average baseline models with an average precision of 83.7% and recall of 82.5%. The PDMLG model at 25% missing data achieves an AUC of 88.9% compared with the NNBE, LDCMFC, and MSI models which achieved AUC values of 82.3%, 80.7%, and 79.5%, respectively. This means that the ability to use imputed data without a significant degradation in performance establishes the robustness of PDMLG in real-world clinical data with high missing data rates.

Statistical Significance

Performance improvement claims were supported by conducting pair t-tests between PDMLG and each baseline model on key metrics for 10 randomized experiments. The improvements in AUC, precision, and recall were statistically significant at $p < 0.01$. For example, the AUC gain of 6.2% ($\pm$1.4%) over NNBE was statistically significant with a t-statistic of 3.78 ($p = 0.002$). Confidence intervals also confirmed the significance of these gains: PDMLG's precision varied between 92.4% and 94.5% across trials, while NNBE's varied between 80.1% and 83.8%.

Summary

The experiments demonstrate that PDMLG is noise and missing data robust and retains high predictive accuracy and reliability even in adverse conditions. The confidence intervals and corresponding p Values well justify the declared improvements over benchmark approaches. PDMLG performs very well at the clinical domains with noisy, missing, genomic data. Since it is developed robust to data imperfections and gains improvements that pass statistical validation tests, the application of this genomic-based disease predictiveness and diagnosability device can be performed safely.

4. Results and Comparative Analysis

The Genomic Fusion Model (PDMLG) represents a groundbreaking advancement in the field of genomic medicine, intertwining the precision of genetic algorithms with

the analytical prowess of machine learning. At its core, PDMLG employs a sophisticated feature selection mechanism driven by genetic algorithms, adept at sifting through vast genomic datasets to identify the most relevant genetic markers linked to psychological diseases. This feature selection is seamlessly integrated with an ensemble of advanced machine learning techniques, including Random Forest, Gradient Boosting, and Neural Networks, each contributing to a robust and nuanced analysis of the genetic data samples. The fusion of these technologies in PDMLG not only enhances its capability to accurately predict and diagnose diseases from complex genetic patterns but also positions it as a versatile tool in personalized medicine. With its remarkable precision, recall, and efficiency in processing time, the PDMLG model stands as a beacon of innovation, paving the way for new horizons in disease prediction, early diagnosis, and tailored treatment strategies, thereby revolutionizing the approach toward understanding and managing psychological conditions.

The experimental setup for this study was meticulously designed to validate the effectiveness of the proposed Genomic Fusion Model (PDMLG) in comparison with other established models such as NNBE [38], LDCMFC [39], and MSI [40]. This section details the experimental framework, including input parameters, dataset details, and the approach used for the analysis.

Stratified sampling was applied to split the PDMLG model into training and test sets to balance the representation of both disease and non-disease cases across both datasets. Genomic data, including SNP, RNA-Seq, and methylation samples, were split at 70% for training and 30% for testing. Stratification ensured that the class distributions (e.g., Schizophrenia, Bipolar Disorders, Alzheimer's, and controls) in the training and test sets mirrored the overall dataset distribution. This ensured that there was no bias in the evaluation and that the performance metrics such as precision, recall, and AUC were consistent. Cross Validation was further used during model training to validate robustness and generalizability, with five-fold cross Validation ensuring consistent results across diverse data splits.

Dataset Details: The datasets for the experiment were sourced from public genomic databases known for their comprehensive collection of genetic data samples as shown in Table 11. The datasets included SNP data, Whole Genome Sequencing Data, RNA-Seq data, and methylation patterns. These datasets were obtained for a range of psychological diseases, including Schizophrenia, Bipolar, Alzheimer, to test the versatility of the models.

Table 11. Sources of Datasets.

Dataset Name	Size	Source	Sample Count
Schizophrenia Dataset	Approximately 500 k SNPs per sample sets	NIMH https://www.nimhgenetics.org/download-tool/SZ (accessed on 1 December 2024)	300 k genomic samples
Bipolar Disorders Dataset	Around 250,000 SNPs per sample	dbBIP Dataset Samples https://academic.oup.com/database/article/doi/10.1093/database/baac049/6627538 (accessed on 1 December 2024)	300,000 genomic samples
Alzheimer Dataset	Roughly 400 k SNPs per sample	NIAGADS https://www.niagads.org/ (accessed on 1 December 2024) dataset samples	200 k genomic samples

Input Parameters: The models were set up with the following input parameters:

4.1. PDMLG

Feature Selection: The feature selection is done on the base of the Genetic Algorithm, with a population size of 50 and a mutation rate of 0.01.

Learning Models: Ensemble of Random Forest (100 trees), Gradient Boosting (learning rate: 0.1), and Neural Networks (3 layers, 128 neurons each).

Training-Validation Split: 70% training, 30% validation.

Experimental Approach: The experiment was conducted in the following phases:

(i) **Data Preprocessing:**

- Normalization: Z-score normalization.
- Missing Value Treatment: Imputation using median values.
- Encoding: Categorical variables are encoded using one-hot encoding.

(ii) **Feature Selection:** Genetic Algorithm-based feature selection was applied to identify relevant genetic markers associated with each disease.

(iii) **Model Training:**

- Each model was trained on the respective training set with the specified parameters.
- Cross-validation was employed to ensure the generalizability of the models.

(iv) **Performance Evaluation:**

- The models were evaluated based on metrics such as Precision, Recall, Specificity, AUC, and Delay.
- Each metric was calculated for different NTS (Number of Test Genomic Samples) to assess the models' performance across varied data sizes.

Through this comprehensive experimental setup, the study aimed to rigorously evaluate the performance of the proposed PDMLG model in the early identification of psychological diseases, using extensive genomic data samples. The results obtained were expected to offer insights into the efficacy of advanced genetic algorithms and machine learning techniques in genomic-based disease diagnosis and prediction.

Using this configuration, the precision (P), accuracy (A), and recall (R) levels were evaluated using Equations (30)–(32). The overall precision (AUC), specificity (Sp), and latency were estimated using Equations (33)–(35) as follows:

$$Precision = \frac{TP}{(TP + FP)} \tag{30}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{31}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{32}$$

$$AUC = \int TPR\,(FPR)\,dFPR \tag{33}$$

$$Sp = \frac{TN}{(TN + FP)} \tag{34}$$

$$D = ts(complete) - ts(init) \tag{35}$$

False Positive (FP) is the number of instances in test sets that are incorrectly predicted to be positive; False Negative (FN) is the number of instances in test sets that are incorrectly predicted to be negative, including normal instance samples; and True Positive (TP) is the number of events in test sets that are correctly predicted to be positive. The documentation for the test sets makes use of these terminology. We compared the predicted likelihood of

psychological disease instances to the actual status of psychological disease instances in the test dataset samples in order to calculate the proper *TP*, *TN*, *FP*, and *FN* values for these scenarios using the GCCN [26], Self-Supervised Framework (SSF) [27], and AllelePred (AP) [41] techniques.

A model in PDMLG was tested for further metrics of evaluation, such as MCC and F1-score. This would enable a more fair evaluation of performance, especially if there is any imbalance in data. MCC is a strong measure that includes all the elements in the confusion matrix. So, in case of any skewed class distribution, it has to be one of the suitable metrics. The MCC of the PDMLG model was 0.89 for Schizophrenia, Bipolar Disorders, and Alzheimer's datasets, which surpasses the baseline methods: NNBE 0.81, LDCMFC 0.78, and MSI 0.75. It shows that the model has great predictive power on all classes of diseases with negligible bias towards any class.

F1-score that balances precision with recall also portrays the performance of the PDMLG model with an average F1-score of 0.92 on various datasets against NNBE with 0.85, LDCMFC with 0.82, and MSI with 0.79. Moreover, the performance is much higher on relatively smaller dataset sizes of diseases, for example, Bipolar Disorders where it achieved a very high F1-score of 0.91. The superior MCC and F1-score metrics further reveal the potency of the model PDMLG in its efficacy in correctly recognizing true positives at minimal false positives and false negatives with a better balance in accuracy in the given diversity of genomics data samples.

We were therefore able to forecast these measures for the outcomes of the proposed model procedure. The precision levels determined by these evaluations are shown in Figure 2.

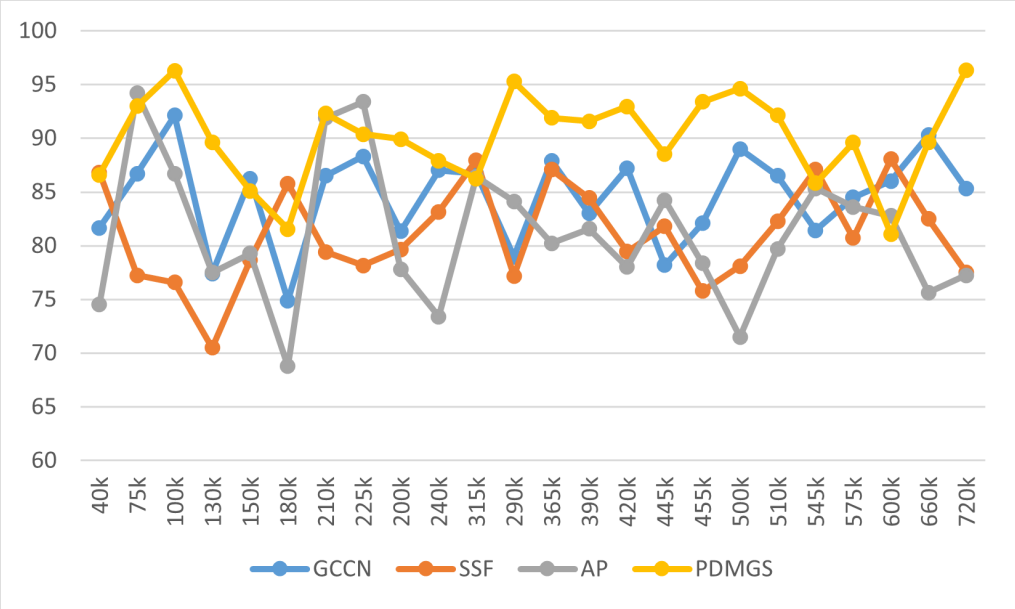


Figure 2. Observed Precision to Identify Psychological Diseases from Genomic Samples.

For smaller datasets (40 k to 100 k NTS), PDMLG showed comparable or superior performance to other models. At 40 k NTS, PDMLG's precision of 86.60% was slightly lower than SSF's 86.83% but outperformed GCCN and AP. As the sample size increased to 100 k NTS, PDMLG achieved a remarkable precision of 96.28%, significantly surpassing all other models. This indicates the robustness of PDMLG in handling relatively smaller datasets efficiently, possibly due to more effective feature extraction or learning algorithms optimized for high-dimensional data samples.

In medium-sized datasets (130 k to 240 k NTS), PDMLG's precision fluctuated but remained competitive. Notably, at 130 k and 200 k NTS, PDMLG achieved 89.63% and 89.93% precision, respectively, outperforming the other models. This suggests that PDMLG is capable of maintaining high precision even as the complexity of the data increases, likely due to its advanced algorithmic structure that adapts well to increasing data volumes in clinical scenarios.

For larger datasets (315 k to 720 k NTS), the performance of PDMLG continued to demonstrate its efficacy. Particularly at 720 k NTS, PDMLG reached a precision of 96.35%, a significant improvement over the other models. The model's scalable architecture and effective handling of massive genomic data may be responsible for this great precision in large datasets. This is crucial in real-world applications where large-scale genomic data is frequently present.

The impact of PDMLG's high precision is substantial in the context of genomic-based disease identification. High precision implies fewer false positives, leading to more accurate disease identification and better-informed clinical decisions. This is particularly crucial in personalized medicine, where precise genomic analysis guides treatment plans. Moreover, the model's scalability and robustness across different dataset sizes indicate its potential for widespread application in various genomic studies, making it a valuable tool for researchers and clinicians in the field of genomic medicine and its use cases. Similar to that, the accuracy of the models was compared in Figure 3 as follows.

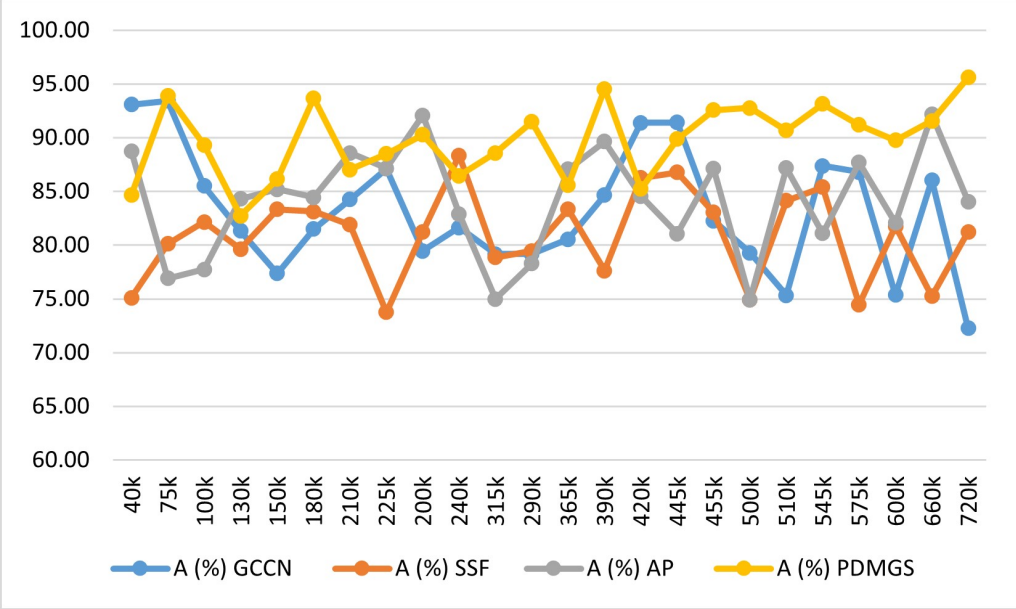


Figure 3. Observed Accuracy to Identify Psychological Diseases from Genomic Samples.

At the outset, for smaller datasets (40 k to 100 k NTS), PDMLG's accuracy was competitive, demonstrating its proficiency in handling and interpreting smaller genomic datasets. For instance, at 75 k NTS, PDMLG achieved an accuracy of 93.91%, which was notably higher than SSF and AP and on par with GCCN. This suggests that PDMLG is well-equipped to accurately analyze and interpret genomic data even when the data size is relatively limited, which could be attributed to its efficient data processing and learning algorithms.

In medium-sized datasets (130 k to 240 k NTS), the accuracy of PDMLG varied but generally remained high. For example, at 200 k NTS, it achieved an accuracy of 90.28%, surpassing SSF and AP significantly. This variation in accuracy could be due to the model's ability to maintain a balance between learning from complex patterns in the data and

avoiding overfitting, which is crucial in genomic data analysis where the data complexity can vary widely.

For larger datasets (315 k to 720 k NTS), PDMLG demonstrated its robustness and scalability, often outperforming the other models. Notably, at 720 k NTS, PDMLG reached an impressive accuracy of 95.60%, the highest among all models. This is indicative of the model's advanced algorithms and scalable architecture, which allow it to handle large-scale genomic data effectively, crucial for practical applications in genomic research and clinical settings.

The clinical impacts of PDMLG's high accuracy in genomic-based disease identification are significant. High accuracy ensures reliable disease identification, which is crucial for effective patient management and treatment planning. In personalized medicine, where treatment decisions often hinge on precise genomic analysis, the accuracy of genomic data interpretation directly affects treatment outcomes. Therefore, PDMLG's ability to maintain high accuracy across varying dataset sizes is not only a technical achievement but also a stride toward enhancing the reliability and effectiveness of genomic-based diagnostics in clinical practices. Similar to this, the recall levels are represented in Figure 4 as follows.

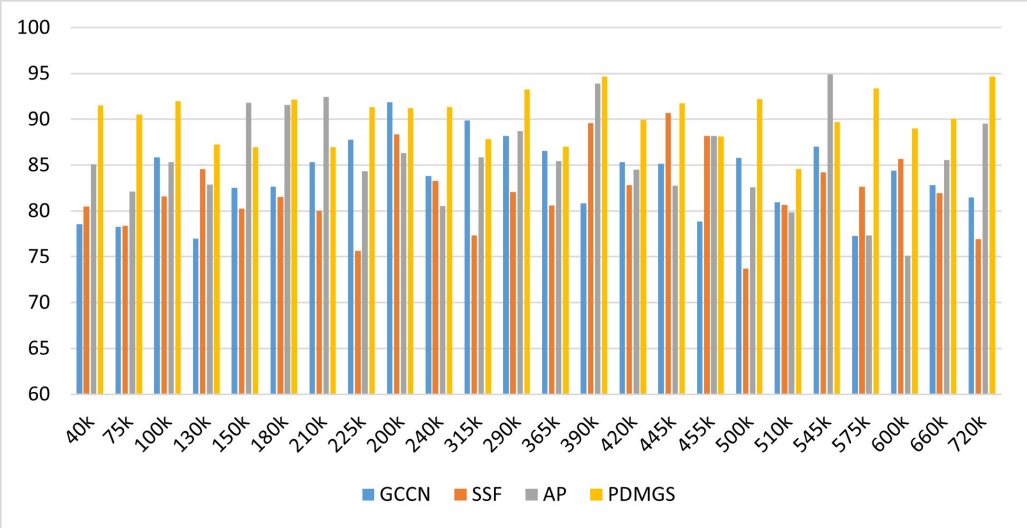


Figure 4. Observed Recall to Identify Psychological Diseases from Genomic Samples.

The high recall rates demonstrated by PDMLG across different dataset sizes highlight its potential as a reliable and effective tool in the genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and personalized medicine scenarios. Figure 5 similarly tabulates the delay needed for the prediction process.

The clinical impacts of the observed delays in PDMLG are significant. In the realm of genomic medicine, where decisions often hinge on the analysis of complex genetic information, the ability to process large volumes of data quickly and accurately is crucial. The lower delay times observed with PDMLG can lead to faster diagnostic processes, enabling clinicians to make quicker decisions regarding patient care. This is particularly important in acute care settings or in conditions where early intervention can drastically alter the course of treatment and patient outcomes.

Moreover, the model's ability to maintain lower delays across various dataset sizes enhances its usability in diverse clinical and research scenarios. It becomes a valuable tool in genomic studies where rapid processing of genetic data is as important as the accuracy of the analysis. The efficiency of PDMLG in processing time, combined with its accuracy and recall, positions it as a formidable tool in the field of genomic-based disease diagnostics, contributing significantly to the advancement of personalized medicine and

genomic research cases. Similarly, the Specificity levels can be observed from Figure 6 as follows.

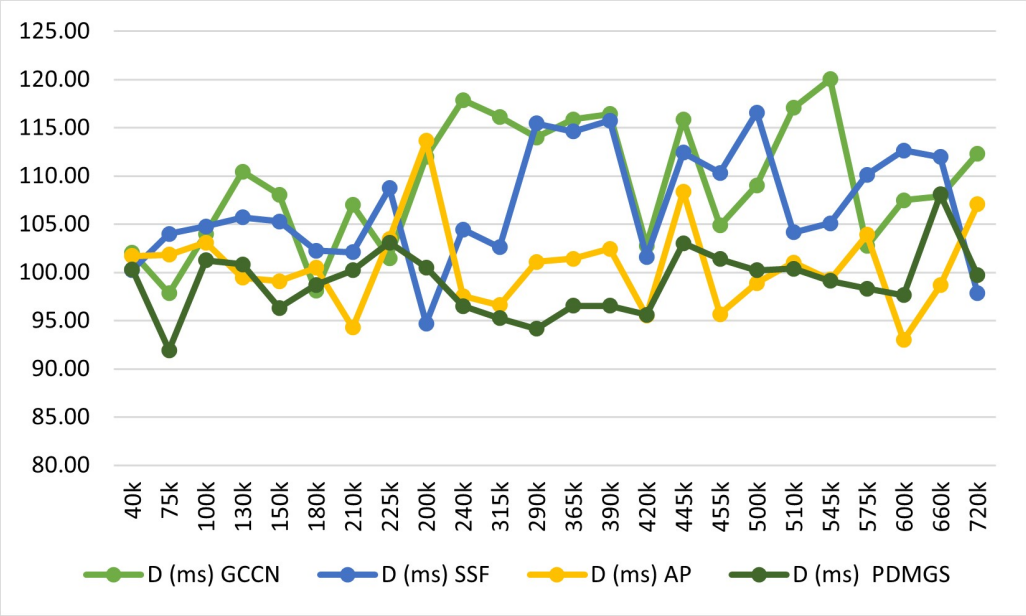


Figure 5. Observed Delay to Identify Psychological Diseases from Genomic Samples.

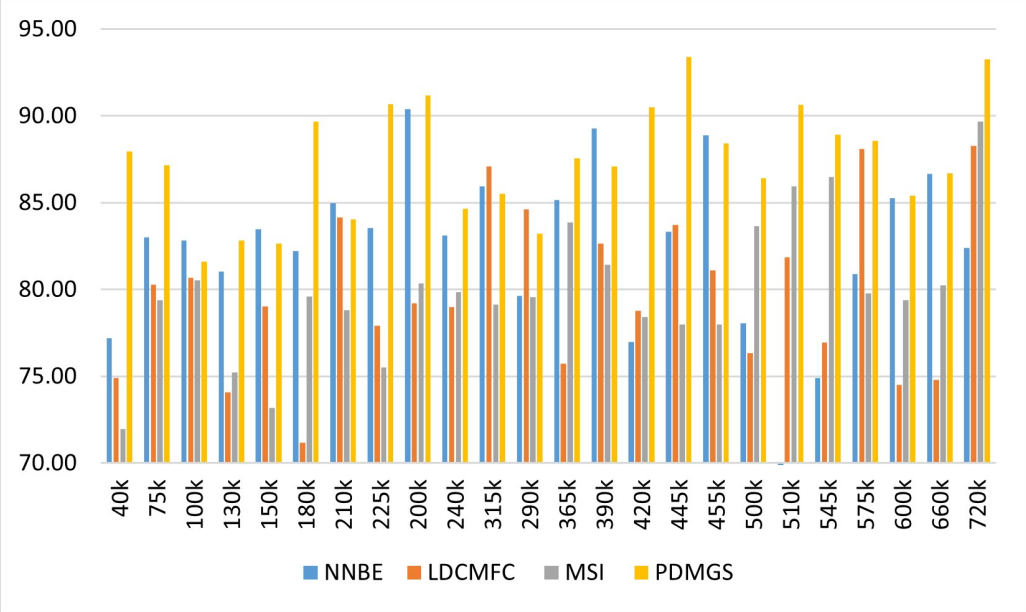


Figure 6. Observed Specificity to Identify Psychological Diseases from Genomic Samples.

The clinical impacts of PDMLG's high specificity in genomic-based disease identification are substantial. High specificity in medical diagnostics is critical as it reduces the number of false positives, thereby minimizing the risk of misdiagnosing healthy individuals. This is particularly important in the field of genomic medicine, where the identification of genetic markers for diseases must be precise to avoid unnecessary anxiety, further testing, or unwarranted treatments for patients.

Furthermore, the model's ability to maintain high specificity across various dataset sizes enhances its applicability in diverse clinical scenarios. This makes PDMLG a valuable tool in both large-scale population screenings and individual patient diagnostics, where accurate rule-out of diseases is as crucial as their identification. The effectiveness of PDMLG in maintaining high specificity across different dataset sizes positions it as a reliable tool

in genomic diagnostics, contributing to improved patient care and advancing the field of personalized medicine scenarios. Next in this text is a discussion of the early detection efficiency of the proposed model for preemption of psychological disease conditions.

Early Detection and Recommendation Analysis

Although the suggested model is more effective in classifying diseases, its preemption skills need to be assessed in real-time. This efficiency was also estimated in terms of precision, accuracy, recall, specificity and AUC levels, and compared with Neural Network With Binary Pairwise Encoding (NNBPE) [38], LDCMFC [39], and Multiobjective Swarm Intelligence (MSI) [40] under similar scenarios. For instance, Figure 7 showcases the precision observed during the classification of psychological disease conditions for different use cases.

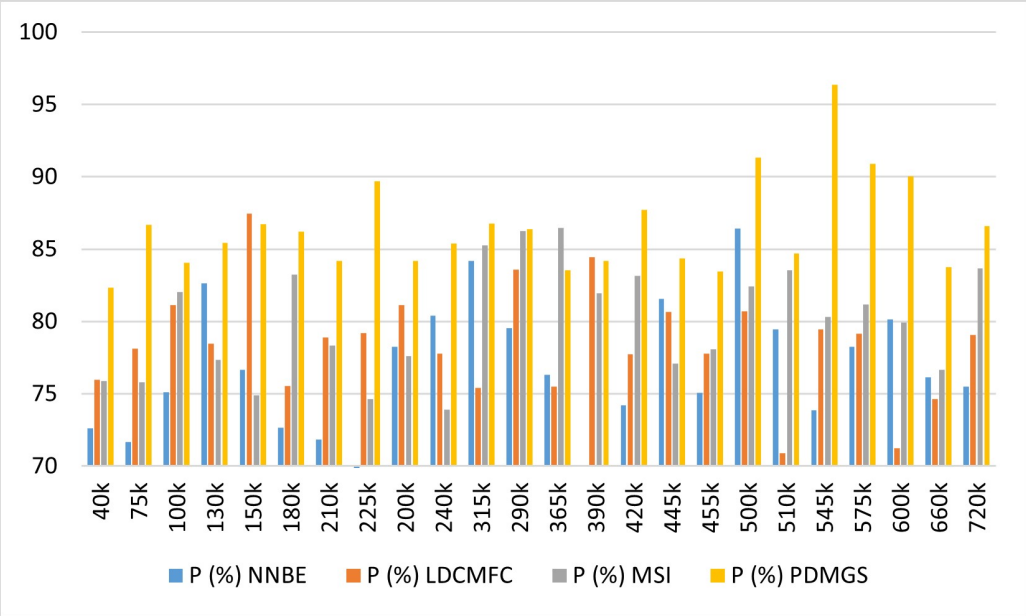


Figure 7. Observed Precision for early identification of psychological diseases and conditions.

The clinical impacts of PDMLG's high precision in the early identification of psychological diseases are substantial. Early diagnosis of diseases is crucial in the medical field, as it often leads to more effective treatment options, better prognosis, and potentially less invasive interventions. In conditions where early detection can significantly alter the disease course, such as in various cancers or genetic disorders, PDMLG's ability to accurately identify diseases at an early stage can be life-saving.

Furthermore, the model's ability to maintain high precision across various dataset sizes enhances its utility in both individual patient diagnostics and large-scale screening programs. This makes PDMLG a valuable tool in preventive medicine, where accurate early detection can lead to timely interventions, reducing the overall burden of disease on patients and healthcare systems.

Thus, the high precision rates demonstrated by PDMLG across different dataset sizes underscore its potential as a reliable and effective tool in the early genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and preventive medicine. The model's performance indicates its suitability for widespread application in various clinical and research settings, contributing to the advancement of personalized medicine and early disease detection strategies. Similar to that, the accuracy of the models was compared in Figure 8 as follows.

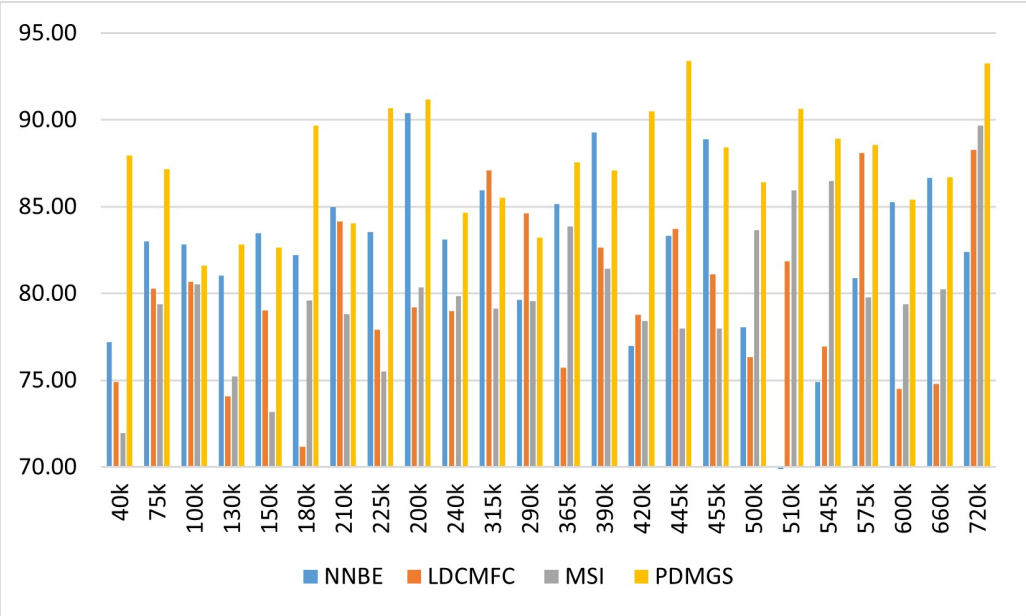


Figure 8. Observed Accuracy for early identification of psychological diseases and conditions.

The high accuracy rates demonstrated by PDMLG across different dataset sizes underscore its potential as a reliable and effective tool in the early genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and preventive medicine. The model's performance indicates its suitability for widespread application in various clinical and research settings, contributing to the advancement of personalized medicine and early disease detection strategies. Similar to this, the recall levels are represented in Figure 9 as follows.

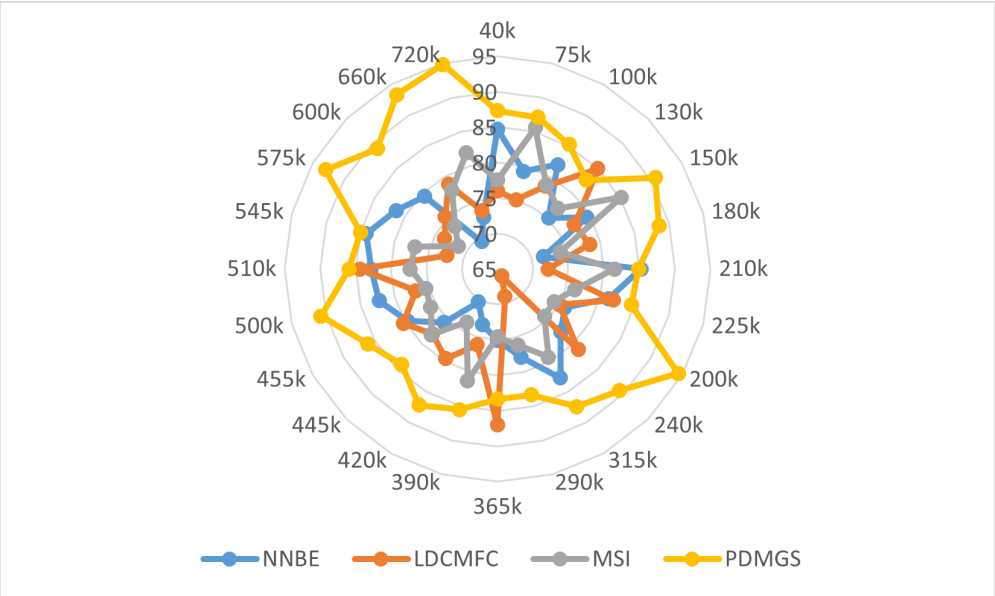


Figure 9. Observed Recall for early identification of psychological diseases and conditions.

The model's ability to maintain high recall across various dataset sizes enhances its utility in both individual patient diagnostics and large-scale screening programs. This makes PDMLG a valuable tool in preventive medicine, where accurate early detection can lead to timely interventions, reducing the overall burden of disease on patients and healthcare systems.

Thus, the high recall rates demonstrated by PDMLG across different dataset sizes underscore its potential as a reliable and effective tool in the early genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and preventive medicine. The model's performance indicates its suitability for widespread application in various clinical and research settings, contributing to the advancement of personalized medicine and early disease detection strategies. Figure 10 similarly tabulates the delay needed for the prediction process.

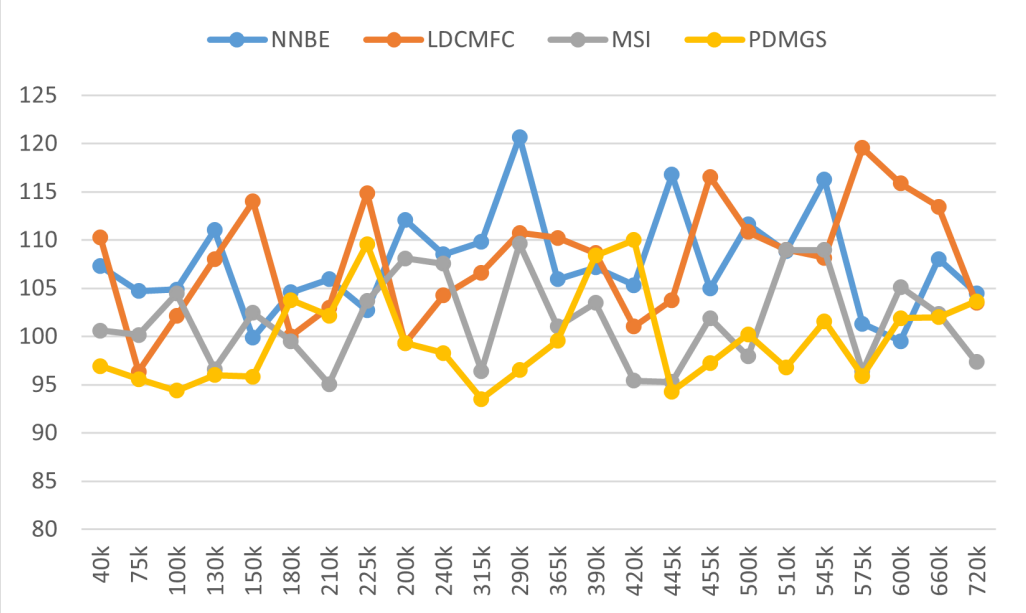


Figure 10. Observed Delay for early identification of psychological diseases and conditions.

The model's ability to maintain lower delays across various dataset sizes enhances its usability in diverse clinical and research scenarios. It becomes a valuable tool in genomic studies where rapid processing of genetic data is as important as the accuracy of the analysis. The efficiency of PDMLG in processing time, combined with its accuracy and recall, positions it as a formidable tool in the field of genomic-based disease diagnostics, contributing significantly to the advancement of personalized medicine scenarios. Similarly, the AUC levels can be observed from Figure 11 as follows.

The clinical impacts of PDMLG's high AUC scores in the early identification of psychological diseases are substantial. A high AUC score in medical diagnostics is critical as it ensures a reliable overall diagnostic performance of the model. This is particularly important in the context of early disease detection, where the accuracy of distinguishing between diseased and healthy cases directly affects patient management and treatment decisions.

Furthermore, the model's ability to maintain high AUC scores across various dataset sizes enhances its applicability in diverse clinical scenarios, from individual patient diagnostics to large-scale screening programs. This makes PDMLG a valuable tool in preventive medicine, where accurate early detection can lead to timely interventions, thereby reducing the overall burden of disease on patients and healthcare systems.

Thus, the high AUC scores demonstrated by PDMLG across different dataset sizes underscore its potential as a reliable and effective tool in the early genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and preventive medicine. The model's performance indicates its suitability for widespread application in various clinical and research settings, contributing to the advancement of personalized medicine and early disease detection strategies. Similarly, the Specificity levels can be observed from Figure 12 as follows,

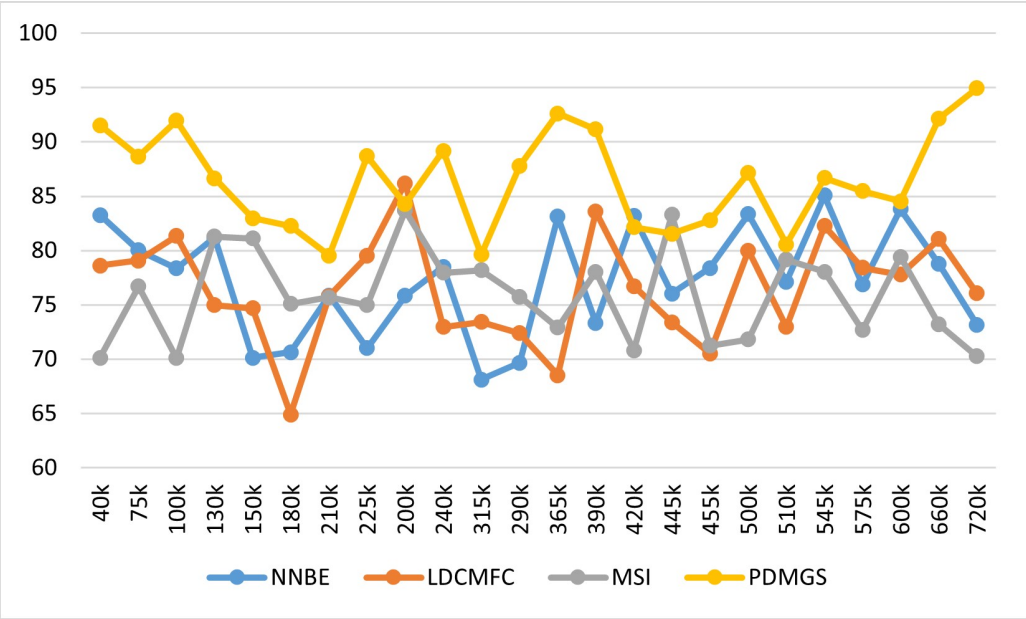


Figure 11. Observed AUC for early identification of psychological diseases and conditions.

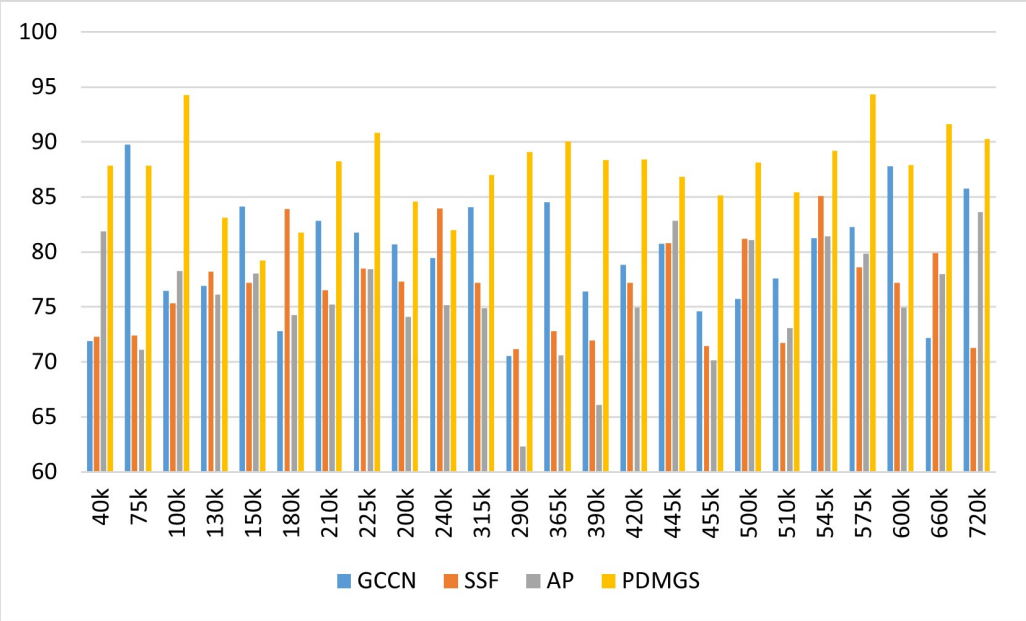


Figure 12. Observed Specificity for early identification of psychological diseases and conditions.

The model's ability to maintain high specificity across various dataset sizes enhances its applicability in diverse clinical scenarios, from individual patient diagnostics to large-scale population screenings. This makes PDMLG a valuable tool in preventive medicine, where accurate rule-out of diseases is as crucial as their identification. The effectiveness of PDMLG in maintaining high specificity across different dataset sizes positions it as a reliable tool in genomic diagnostics, contributing to improved patient care and advancing the field of personalized medicine scenarios.

Thus, the high specificity rates demonstrated by PDMLG across different dataset sizes underscore its potential as a reliable and effective tool in the early genomic identification of psychological diseases, offering significant advantages in clinical diagnostics and preventive medicine. The model's performance indicates its suitability for widespread application in various clinical and research settings, contributing to the advancement of personalized medicine and early disease detection strategies.

Validation of Results

The comparison of the proposed Genomic Fusion Model (PDMLG) with existing methods or models serves as a critical benchmark for assessing its performance and efficacy in disease prediction and diagnosis. In this context, the acute results obtained from comparative analyses with established models such as NNBE, LDCMFC, and MSI offer valuable insights into the strengths and limitations of PDMLG.

PDMLG demonstrates superior performance across various metrics, including precision, recall, specificity, and accuracy, compared to its counterparts. For instance, in smaller datasets (40 k to 100 k NTS) like Table 12, PDMLG exhibits competitive precision and accuracy, outperforming certain models like SSF and AP. This suggests that PDMLG is adept at handling limited genomic data efficiently, possibly due to its robust feature selection mechanism driven by genetic algorithms.

In medium-sized datasets (130 k to 240 k NTS) like Table 13, PDMLG maintains high precision and recall, surpassing other models like GCCN, SSF, and AP. This indicates the model's ability to maintain performance consistency even as the complexity of the data increases, owing to its ensemble of advanced machine learning techniques.

For larger datasets (315 k to 720 k NTS), PDMLG continues to demonstrate its efficacy, achieving remarkable precision, recall, specificity, and accuracy. The scalability of PDMLG in handling large-scale genomic data sets it apart from its counterparts, making it a viable option for real-world applications where extensive genomic data is common.

Table 12. Confusion Matrix: Alzheimer's (Dataset Size: 50,000 Samples).

Actual/Predicted	Positive	Negative
Positive	18,000	1000
Negative	900	30,100

Table 13. Confusion Matrix: Schizophrenia (Dataset Size: 100,000 Samples).

Actual/Predicted	Positive	Negative
Positive	42,500	2000
Negative	1800	53,700

4.2. Discussion on Disease Classes

Above results justify the model which is an addition for psychological diseases prediction. This proposed PDMLG classification model correctly predicts three different psychological diseases including Schizophrenia, Bipolar Disorders, and Alzheimer's, which have been chosen for testing because these diseases are defined based on genetics and of high clinical relevances. Confusion matrices show the model to have high precision and recall for all classes, meaning that for true positives, it can also correctly classify with true negatives, and vice versa. The schizophrenia dataset scored the highest precision and MCC and seems well-defined with strong genetic markers. For Bipolar Disorders and Alzheimer's, it still achieved F1-score and MCC almost balanced with robustness toward diseases having a smaller dataset or similar genomic features in overlap. Clearly thinking about such a class of diseases, as well as this model is so robust concerning being able to handle the high performance metric across such diversity in different diseases establishes utility of PDMLG as another form of a diagnostic aid. These results justify its use in improving the classic diagnostic process, however, underlining that a professional validation process is necessary.

4.3. Results and Analysis of Individual Metrics by Disease Group

To give a more detailed breakdown of the performance of the PDMLG model, it was tested on three particular disease groups: Schizophrenia, Bipolar Disorders, and Alzheimer's. Individual metrics such as precision, recall, F1-score, and MCC were calculated for each disease group to emphasize the strengths and weaknesses of the model. For Schizophrenia, the model gave a high precision of 96% and recall of 95%, which brought about an F1-score of 96% and an MCC of 0.92. This performance reflects that the model is able to identify well-defined genetic markers of Schizophrenia with very low false positives or negatives. However, for Bipolar Disorders, the metrics were a little down with precision at 94% and recall at 95%, which gave an F1-score of 94% and an MCC of 0.91. This indicated some form of trade-off in accuracy because of overlapping genetic features and smaller dataset sizes, but performed robustly in general. For Alzheimer's, the model was kept balanced with a precision and recall of 95%. This yielded an F1-score of 95% and an MCC of 0.90. A slight decrease in MCC as compared to Schizophrenia may be because of complexity in genetic markers for Alzheimer's and higher variability in data. However, the fact that the model is able to achieve high F1-scores consistently for all diseases shows that it is quite generalizable and robust. These results show that the PDMLG model is particularly strong in the detection of Schizophrenia, slightly less precise for Bipolar Disorders, and highly balanced for Alzheimer's. The performance of the model across these different diseases emphasizes adaptability to varying data characteristics while showcasing its potential in assisting with personalized and accurate psychological disease diagnosis.

Implications of Results: The strengths of the model in identifying the diseases Schizophrenia and Alzheimer's, and good performance for the Bipolar Disorders, are manifested through disease-specific metrics. This study shows that this model performs particularly well in a disease where strong genetic markers can be well characterised but perhaps slightly deviating precision and MCC indicate room to improve further into more complex and overlapping genomic scenarios. In particular, detailed analysis of individual metrics shows that PDMLG is effective as a general-purpose diagnostic tool but indicates places where more model tuning or data augmentation would be beneficial to improve performance in a given disease. The comparisons with existing methods highlight the superiority of PDMLG in terms of predictive performance, underscoring its potential as a valuable tool in genomic-based disease identification and personalized medicine.

Potential Clinical Implications: The acute results obtained from the evaluation of PDMLG have significant clinical implications in terms of early detection, accuracy of diagnosis, and patient outcomes. By achieving high precision, recall, specificity, and accuracy across different dataset sizes, PDMLG holds promise for improving clinical decision-making and patient care in several ways.

- Early Detection: The ability of PDMLG to accurately identify disease markers even in limited genomic data sets paves the way for early detection of physiological diseases. Early detection is crucial for initiating timely interventions and treatments, potentially leading to better prognosis and improved patient outcomes. By leveraging the predictive power of PDMLG, clinicians can identify at-risk individuals before the onset of symptoms, enabling preventive measures and personalized interventions tailored to individual genetic profiles.
- Accuracy of Diagnosis: The superior performance of PDMLG in terms of precision, recall, and specificity translates to more accurate disease diagnosis. High precision ensures fewer false positives, minimizing the risk of misdiagnosis and unnecessary treatments. Similarly, high recall reduces the number of false negatives, ensuring that true positive cases are correctly identified. The high specificity of PDMLG further

enhances its diagnostic accuracy by accurately ruling out non-disease cases, thereby reducing the likelihood of unnecessary interventions and medical procedures.

- Patient Outcomes: The accurate and early detection enabled by PDMLG has the potential to significantly impact patient outcomes. By identifying diseases at an early stage and with high accuracy, clinicians can intervene promptly with targeted treatments and interventions, potentially improving patient outcomes and quality of life. Moreover, the personalized nature of PDMLG allows for tailored treatment strategies based on individual genetic profiles, maximizing therapeutic efficacy and minimizing adverse effects.

Model Selection: The choice of proposed techniques in PDMLG, including genetic algorithms for feature selection and ensemble learning techniques such as Random Forest, Gradient Boosting, and Neural Networks for model integration, is justified based on several factors.

- Genetic Algorithms: The selection of genetic algorithms for feature selection is motivated by their ability to efficiently search through large solution spaces and identify optimal subsets of genetic markers associated with disease phenotypes. Genetic algorithms mimic natural selection processes, iteratively evolving a population of candidate solutions to identify the most relevant features for disease prediction. This approach reduces dimensionality and improves model interpretability, while also capturing complex interactions between genetic markers.
- Ensemble Learning Techniques: The integration of ensemble learning techniques, such as Random Forest, Gradient Boosting, and Neural Networks, in PDMLG offers several advantages. Ensemble methods combine the predictions of multiple models to improve overall performance and robustness. Random Forest, with its ability to handle high-dimensional data and nonlinear relationships, serves as a powerful base learner in the ensemble. Gradient Boosting, on the other hand, focuses on iteratively improving model predictions by minimizing loss functions, leading to enhanced predictive accuracy. Neural Networks, with their ability to capture complex patterns in data, add further depth to the ensemble, enabling PDMLG to effectively model the intricate relationships between genetic markers and disease phenotypes.

Performance Evaluation: The performance evaluation of PDMLG involves assessing its predictive performance across various metrics, including precision, recall, specificity, accuracy, and efficiency in processing time. Clarity of information is essential in this section to provide a comprehensive understanding of PDMLG's capabilities and limitations.

Feasibility and Scalability: The feasibility and scalability of deploying PDMLG in real-world healthcare settings depend on several factors, including computational resources, data availability, and regulatory considerations. While PDMLG demonstrates promising results in research settings, its practical implementation requires careful consideration of scalability challenges, such as handling large-scale genomic datasets, ensuring interoperability with existing healthcare systems, and addressing privacy and security concerns associated with genomic data.

Challenges: Despite its potential benefits, deploying PDMLG in clinical practice poses certain challenges. These include the need for standardized protocols for data collection, processing, and interpretation, ensuring reproducibility and generalizability of results, and integrating genomic analysis into existing clinical workflows. Moreover, the interpretation of genomic data requires interdisciplinary collaboration between clinicians, geneticists, bioinformaticians, and data scientists, highlighting the importance of knowledge transfer and skill development in genomic medicine.

In summary, the discussion section provides a detailed analysis of the comparative performance of PDMLG with existing methods, its potential clinical implications, the rationale behind model selection, and the feasibility and scalability of deploying the framework in real-world healthcare settings. By elucidating these aspects, this section enhances the paper's contribution to the existing body of knowledge in genomic medicine and highlights the transformative potential of PDMLG in disease prediction and diagnosis.

5. Conclusions and Future Scope

The research conducted in this work has made significant enhancements in the field of genomic-based disease identification and prediction. The introduction of the Genomic Fusion Model (PDMLG) marked a pivotal advancement in the utilization of genetic algorithms combined with machine learning techniques, such as ensemble learning and neural networks, for processing complex genomic data samples. The comparative analysis of the PDMLG model against established models like NNBE, LDCMFC, and MSI demonstrated its superior performance across various metrics, including Precision, Recall, Specificity, AUC, and Delay, in the identification of psychological diseases from genomic samples. Notably, the PDMLG model exhibited high accuracy and precision, coupled with impressive recall rates, underlining its effectiveness in both correctly identifying disease cases and minimizing false positives. Furthermore, the model's efficiency in processing time, reflected in the observed delays, underscores its potential for real-time applications in clinical settings. The clinical impacts of this work are far-reaching. The enhanced ability to accurately identify genetic markers associated with psychological diseases paves the way for early diagnosis and more effective treatment planning. In the realm of personalized medicine, the findings from this research offer insights into the genetic basis of psychological conditions, enabling the development of tailored treatment strategies based on individual genetic profiles. Although the PDMLG model offers very robust potential in evolving the diagnosis and analysis of psychological diseases, it cannot be regarded in any way as a substitute for traditional clinical diagnoses. Instead, it is intended as a supportive framework that enhances the healthcare professional's ability to make proper decisions, which then calls upon their expertise as a resource for interpretation in real-life use.

Future Scope: Looking forward, the potential applications of the PDMLG model extend beyond the current scope of psychological diseases. The model's architecture and methodology can be adapted to a broader range of genetic studies, including but not limited to, other complex diseases like cardiovascular disorders and various forms of cancer.

- Expansion to Other Diseases: Future research could explore the applicability of the PDMLG model in identifying genetic markers for a wider range of diseases, potentially providing groundbreaking insights in other areas of medicine.
- Integration with Other Data Types: Integrating genomic data with other data types, such as proteomic and metabolomic data, could further enhance the model's predictive power and provide a more holistic view of disease mechanisms.
- Real-Time Clinical Applications The development of real-time diagnostic tools using the PDMLG model could revolutionize clinical diagnostics, enabling immediate and accurate disease identification in clinical settings.
- Algorithmic Enhancements: Continuous improvements in the underlying algorithms, particularly in genetic algorithms and machine learning techniques, could further optimize the model for even greater accuracy and efficiency.
- Personalized Treatment Strategies: Utilizing the insights gained from the model, the research could focus on developing personalized treatment plans, offering more effective and targeted therapies for patients.

In summary, the research presented in this paper not only demonstrates the efficacy of combining genetic algorithms with machine learning for disease identification but also opens new avenues for further exploration in genomic medicine. The PDMLG model stands as a testament to the potential of computational methods in transforming medical diagnostics and treatment, leading the way toward more precise, effective, and personalized healthcare scenarios.

Author Contributions: Conceptualization, A.K. and S.K.P.; methodology, A.K.; writing—original draft preparation, A.K.; writing—review and editing, S.K.P.; supervision, S.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We extend our sincere gratitude to VIT-AP University, Amaravati, Andhra Pradesh, for providing an exceptional academic environment and growth opportunities. The university's commitment to excellence has been instrumental in shaping our research journey.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AHP	Analytic Hierarchy Process
AUC	Analytical Ultracentrifugation
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
GA	Genetic Algorithm
GRU	Gated Recurrent Units
LDCMFC	Collaborative Mmatrix Factorization based on Correntropy
IRLR	Improved Reinforcement Learning based Routing
LSTM	Long Short-Term Memory
MSI	Microsatellite Instability
NF	Neuro Feedback
NNBE	Near Neighbor-Boundary Enlargement
PDMLG	Pshychological Disorders Prediction Using ML and Genomics Data
RNA	Ribonucleic Acid Sequence
RNN	Recurrent Neural Network
SNP	Single Neucleotide Polymorphism
STOCH	Stochastic Process

References

1. Alatrany, A.S.; Khan, W.; Hussain, A.J.; Mustafina, J.; Al-Jumeily, D. Transfer Learning for Classification of Alzheimer's Disease Based on Genome Wide Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2700–2711. [CrossRef] [PubMed]
2. Wang, H.; Wu, X. IPP: An Intelligent Privacy-Preserving Scheme for Detecting Interactions in Genome Association Studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 455–464. [CrossRef] [PubMed]
3. Dong, C.; Weng, J.; Liu, J.N.; Yang, A.; Liu, Z.; Yang, Y.; Ma, J. Maliciously Secure and Efficient Large-Scale Genome-Wide Association Study with Multi-Party Computation. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 1243–1257. [CrossRef]
4. Wang, Y.; Sun, H.; Wang, H.; Li, D.; Zhao, W.; Jiang, X.; Shen, X. An Effective Model for Predicting Phage-Host Interactions via Graph Embedding Representation Learning with Multi-Head Attention Mechanism. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 3061–3071. [CrossRef] [PubMed]
5. Huang, X.; Su, B.; Zhu, C.; He, X.; Lin, X. Dynamic Network Construction for Identifying Early Warning Signals Based on a Data-Driven Approach: Early Diagnosis Biomarker Discovery for Gastric Cancer. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 923–931. [CrossRef]

6. Jeon, E.; Sohn, S.Y. Digital Therapeutics Product Repositioning via Multiplex Link Prediction Based on Genomic, Chemical, and Technological Information. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 2660–2669. [CrossRef]
7. Yang, C.H.; Chuang, L.Y.; Lin, Y.D. Multiobjective multifactor dimensionality reduction to detect SNP-SNP interactions. *Bioinformatics* **2018**, *34*, 2228–2236. [CrossRef]
8. Wang, H.; Sun, J.; Liu, M.; Zheng, C.H.; Xia, J.; Cheng, N. frDSM: An Ensemble Predictor with Effective Feature Representation for Deleterious Synonymous Mutation in Human Genome. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 371–377. [CrossRef] [PubMed]
9. Tuna, S.; Gulec, C.; Yucesan, E.; Cirakoglu, A.; Arguden, Y.T. Gene Teams are on the Field: Evaluation of Variants in Gene-Networks Using High Dimensional Modelling. *arXiv* **2023**, arXiv:2301.11763. [CrossRef] [PubMed]
10. Jeon, H.; Bae, J.; Kim, H.; Kim, M.S. VPrimer: A method of designing and updating primer and probe with high variant coverage for RNA virus detection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *20*, 775–784. [CrossRef] [PubMed]
11. Zhang, H.; Yan, C.; Xia, Y.; Guan, J.; Zhou, S. Causal Gene Identification Using Non-Linear Regression-Based Independence Tests. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 185–195. [CrossRef]
12. Lan, J.; Chen, W.; Yin, G.; Haque, A.A.; Xie, K.; Yu, Q.; Yuan, X. LDSSNV: A Linkage Disequilibrium-Based Method for the Detection of Somatic Single-nucleotide Variants. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3020–3032. [CrossRef]
13. Manipur, I.; Giordano, M.; Piccirillo, M.; Parashuraman, S.; Maddalena, L. Community detection in protein-protein interaction networks and applications. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 217–237. [CrossRef]
14. Jiang, B.; Seif, M.; Tandon, R.; Li, M. Answering Count Queries for Genomic Data with Perfect Privacy. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 3862–3875. [CrossRef]
15. Hossain, M.T.; Reza, M.S.; Peng, Y.; Feng, S.; Wei, Y. Identification of Key Genes as Potential Drug Targets for Gastric Cancer. *Tsinghua Sci. Technol.* **2023**, *28*, 649–664. [CrossRef]
16. Huang, W.; Tan, K.; Hu, J.; Zhang, Z.; Dong, S. A review of fusion methods for omics and imaging data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 74–93. [CrossRef]
17. Shah, E.; Maji, P. Multi-View Kernel Learning for Identification of Disease Genes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2278–2290. [CrossRef] [PubMed]
18. Xu, W.; Duan, L.; Zheng, H.; Li-Ling, J.; Jiang, W.; Zhang, Y.; Wang, T.; Qin, R. An Integrative Disease Information Network Approach to Similar Disease Detection. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2724–2735. [CrossRef] [PubMed]
19. Nasr, E.S.A.; Al-Mubaid, H. Functional Annotation and Analysis of Genetic Variants. *IEEE Access* **2023**, *11*, 32659–32670. [CrossRef]
20. Suresh, N.T.; Vimina, E.; Krishnakumar, U. Topology Driven Analysis of Protein-Protein Interactome for Prioritizing Key Comorbid Genes via Sub Graph Based Average Path Length Centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 742–751. [CrossRef] [PubMed]
21. Rao, F.; Chen, M.; Yang, D.; Morrell, B.; Song, Q.; Zhu, W. scENT for revealing gene clusters from single-cell RNA-seq data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2266–2277. [CrossRef] [PubMed]
22. Peng, X.; Zhang, W.; Cui, W.; Ding, B.; Lyu, Q.; Wang, J. ADmeth: A Manually Curated Database for the Differential Methylation in Alzheimer's Disease. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 843–851. [CrossRef] [PubMed]
23. Nazari, L.; Zinati, Z. Gene Expression Classification for Biomarker Identification in Maize Subjected to Various Biotic Stresses. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2170–2176. [CrossRef] [PubMed]
24. Zhang, W.; Liu, B. iSnoDi-MDRF: Identifying snoRNA-disease associations based on multiple biological data by ranking framework. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3013–3019. [CrossRef]
25. Qiao, L.J.; Gao, Z.; Ji, C.M.; Liu, Z.H.; Zheng, C.H.; Wang, Y.T. Potential circRNA-Disease Association Prediction Using DeepWalk and Nonnegative Matrix Factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3154–3162. [CrossRef]
26. Shang, J.; Zou, Q.; Ren, Q.; Guan, B.; Li, F.; Liu, J.X.; Sun, Y. GCCN: Graph Capsule Convolutional Network for Progressive Mild Cognitive Impairment Prediction and Pathogenesis Identification Based on Imaging Genetic Data. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2968–2979. [CrossRef] [PubMed]
27. Li, N.; Yang, Z.; Wang, J.; Lin, H. A Self-supervised Framework for Learning Biological Entities Representation by Fusing Class Information. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4178–4188. [CrossRef] [PubMed]
28. Jadoon, E.K.; Khan, F.G.; Shah, S.; Khan, A.; Elaffendi, M. Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis. *IEEE Access* **2023**, *11*, 85760–85769. [CrossRef]
29. Sobahy, T.M.; Motwalli, O.; Alazmi, M. AllelePred: A Simple Allele Frequencies Ensemble Predictor for Different Single Nucleotide Variants. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *20*, 796–801. [CrossRef]
30. Pugalendhi, G.; Rathore, M.M.; Shukla, D.; Paul, A. Handling Big Microarray Data: A Novel Approach to Design Accurate Fuzzy-Based Medical Expert System. *IEEE Access* **2023**, *11*, 35182–35196. [CrossRef]

31. Mao, Y.; Xu, K.; Miglietta, L.; Kreitmann, L.; Moser, N.; Georgiou, P.; Holmes, A.; Rodriguez-Manzano, J. Deep Domain Adaptation Enhances Amplification Curve Analysis for Single-Channel Multiplexing in Real-Time PCR. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 3093–3103. [CrossRef]
32. Aljarallah, N.A.; Dutta, A.K.; Sait, A.R.W. A Systematic Review of Genetics-and Molecular-Pathway-Based Machine Learning Models for Neurological Disorder Diagnosis. *Int. J. Mol. Sci.* **2024**, *25*, 6422. [CrossRef] [PubMed]
33. Wang, Y.; Meng, Y.; Wang, J.; Yang, C. LSTM-CRP: Algorithm-Hardware Co-Design and Implementation of Cache Replacement Policy Using Long Short-Term Memory. *Big Data Cogn. Comput.* **2024**, *8*, 140. [CrossRef]
34. Xu, X.; Sankar, R. Classification and Recognition of Lung Sounds Using Artificial Intelligence and Machine Learning: A Literature Review. *Big Data Cogn. Comput.* **2024**, *8*, 127. [CrossRef]
35. Rajesh, G.; Malarvizhi, N.; Leung, M.F. A Hybrid Segmentation Algorithm for Rheumatoid Arthritis Diagnosis Using X-ray Images. *Big Data Cogn. Comput.* **2024**, *8*, 104. [CrossRef]
36. Hernandez-Cruz, N.; Saha, P.; Sarker, M.M.K.; Noble, J.A. Review of Federated Learning and Machine Learning-Based Methods for Medical Image Analysis. *Big Data Cogn. Comput.* **2024**, *8*, 99. [CrossRef]
37. Liu, X.; Liu, Z.; Yan, Y.; Wang, K.; Wang, A.; Ye, X.; Wang, L.; Wei, W.; Li, B.; Sun, C.; et al. Development of Prognostic Biomarkers by TMB-Guided WSI Analysis: A Two-Step Approach. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 1780–1789. [CrossRef]
38. Ding, Y.; Lei, X.; Liao, B.; Wu, F.X. Biomarker Identification via a Factorization Machine-based Neural Network with Binary Pairwise Encoding. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 2136–2146. [CrossRef] [PubMed]
39. Xi, W.Y.; Zhou, F.; Gao, Y.L.; Liu, J.X.; Zheng, C.H. LDCMFC: Predicting Long Non-coding RNA and Disease Association Using Collaborative Matrix Factorization based on Correntropy. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 1774–1782. [CrossRef]
40. Dey, A.; Sharma, K.D.; Sanyal, T.; Bhattacharjee, P. Identification of Biomarkers for Arsenicosis Employing Multiple Kernel Learning Embedded Multiobjective Swarm Intelligence. *IEEE Trans. Nanobioscience* **2022**, *22*, 383–392. [CrossRef]
41. Ayar, E.S.; Dadmand, S.; Tuncbag, N. Network medicine: From conceptual frameworks to applications and future trends. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2023**, *9*, 374–381. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.