Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 17s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Disease Prediction using Gene Data Over Data Mining and Artificial Intelligence Techniques: A Survey

Paparayudu.nagara[1], Dr.D. Ramesh[2]

[1]Assistant Professor in Information Technology,TKR College of Engineering and Technology,Hyderabad

[2]Professor in Computer Science and Engineering,JNTU,Palair,Khammam,Telangana

Cite this paper as: Paparayudu.nagara, Dr.D. Ramesh, (2025) Disease Prediction using Gene Data Over Data Mining and Artificial Intelligence Techniques: A Survey. *Journal of Neonatal Surgery*, 14 (17s), 861-871.

ABSTRACT

Medical research has investigated Disease Prediction (DP) based on Gene data to a key level today, where the DP is attained by using Data Mining (DM) and Artificial Intelligence (AI) to detect disease-related genes. The traditional methods, such as Genome-Wide Association Studies (GWAS) and Linkage Analysis (LA), typically generate several positional candidate genes; experimental validation is cost-effective and in a time frame. Once Gene Prioritization (GP) methods have been included in computational means such as Feature Selection (FS), clustering, and Machine Learning (ML), GP has been significantly enhanced. Using Deep Learning (DL), Support Vector Machines (SVM), or ensemble classifiers, AI supports the improvement of predicting accuracy based on the learning of fine-grain patterns from genomic datasets. Network-based methods such as protein-protein interaction (PPI) networks and gene ontology (GO) analysis help us to recognize disease-gene predictions (DGP). Next Generation Sequencing (NGS) presents massive genomic data subject to efficient pre-processing and dimensionality reduction methods to mitigate high-dimensionality problems. The DL is used to retrieve the hidden relationships in genomic databases that, in turn, help toward the early disease diagnosis. The developments in integrating heterogeneous genomic data and dealing with biases in training datasets have not yet been attained. This analysis classifies computational tools for gene DP in terms of conceptual model rather than technical method and presents recent works in AI-based genomic research proof towards precision medicine and personal healthcare.

Keywords: *Gene data, accuracy, machine learning, healthcare, disease prediction.*

1. INTRODUCTION

Human Genetic Disorder (GO) requires identifying a low number of 'disease-associated' genes. GD causing many diseases has been detected primarily by traditional methods, including Genome-Wide Association Studies (GWAS), Linkage Analysis (LA), and Positional Cloning (PC). The detection methods generally result in 100 to 1000 candidate genes; experimental proof of these is cost-expensive and time frame [1]. The genomic databases are generated due to the development of Next Generation Sequencing (NGS), a faster and computationally efficient method of training to prioritize the candidate genes, which has become more and more critical [2]. Artificial Intelligence (AI) and Data Mining (DM) are incorporated into—Disease-Gene Prediction (DGP), which has been initiated to be a promising solution to reduce the time taken and enhance the accuracy of the gene-disease association studies.

Machine Learning (ML), Deep Learning (DL), and statistical modeling-based computational methods based on AI deployed on complex genomic data. Such methods can determine disease-relevant patterns, correlations, and biomarkers for early DP and precision medicine. The DM for extracting helpful visions from high-scale genomic datasets is used for further investigation [3]. AI-based models have been proven to be required to method high-dimensional datasets, and the ability to do that has made it an integral part of Gene Prioritization (GP). In the traditional methods, where one is frequently required to define text errors upon and claim prior hypotheses, AI can figure out from the data autonomously apply on earlier never experimental genomic types.

Several computational strategies have tackled disease-gene prediction (DGP) and can be divided into network-based, ML, and functional annotation-based approaches. To infer relationships between genes and diseases, network-based methods analyze biological networks, *e.g.*, Protein-Protein Interaction (PPI) networks and gene expression networks based on specific data on DGP. The basis of these methods is that functionally related genes are likely to be clustered in biological networks and that disease-related genes have network connectivity patterns [4]. The ML uses supervised, unsupervised, and semi-supervised learning methods to assign genes into classes based on different genomic and transcriptomic features [5]. Classifying disease and genes with Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN) have been extensively used to outperform traditional statistical models. Functional annotation-based methods rely on

biological databases such as Gene Cards, Online Mendelian Inheritance in Man (OMIM), and Kyoto Encyclopedia of Genes and Genomes (KEGG) to annotate the candidate genes using the known disease-related functions and pathways to perform Gene Prioritization (GP) [6]. AI and DM science have been able to apply techniques to DGP, which has brought many developments [7]. DL has predicted the non-coding regulatory elements linked to GD. The Transfer Learning (TL) methods have further enhanced the predictive power of AI by using pre-trained models on diverse genomic datasets. Federated Learning (FL) models have also appeared to solve data privacy problems by supporting the collaborative analysis of genomic data across multiple resources without compromising patient privacy [8].

These advancements do not solve the challenge of incorporating heterogeneous genomic data, removing biases in training data, nor provide the interpretability of AI-drive predictions [9]. The standard genomic datasets suffer from a class imbalance that makes disease-associated genes underrepresented compared to non-disease genes, which is why the generalizability of ML is valuable. The black-box nature of DL advances problems regarding model interpretability, demanding the development of Explainable AI (XAI) in genomic research. Future developments in DGP will likely involve multi-modal data integration, incorporating genomic, transcriptomic, epigenetic, and medical data to enhance predictive accuracy. Adopting hybrid models combining knowledge-driven and data-driven methods will further refine candidate GP [10].

This paper presents a comprehensive analysis of computational methods for DGP, focusing on theoretical methodologies rather than technical details. This paper classifies existing bioinformatics tools based on their underlying principles and highlights their help to DGP. This paper discourses the advantages and limitations of AI-based models in genomic research and outlines potential future directions in this evolving field. By synthesizing current advancements and challenges, this review aims to provide valuable visions into the role of AI and DM in accelerating DGP using gene data.

2. IMPACT OF AI AND DATA MINING TECHNIQUES IN DISEASE PREDICTION USING GENES

With AI and DM's integration for gene sequences in DP, genomic analysis's accuracy, efficiency, and scalability have increased. Typically, many genes generated with traditional methods, such as GWAS and LA, are cost-effective, time frame, and experimentally validated. The ML, DL, and bioinformatics methods are becoming the new precursor in pioneers capable of resulting in DGP using the classification of complicated patterns, associations, and mutations in GD.

DM has enabled feature extraction (FE) from high-dimensional genomic datasets, such as clustering, classification, and association rule mining. Dimensionality reduction methods using feature selection (FS) are used to reduce the dimensional complexity of the data so that prediction models become more interpretable and efficient. Several MLs, such as SVM, RF, and DNN, are used to classify genes based on sequence variations, whereas Natural Language Processing (NLP) is used to improve predictions by analyzing genomic literature. Functional annotation-based methods leverage biological databases to refine gene-disease associations.

Network-based methods analyze PPI networks, gene co-expression, and pathway enhancement to start disease correlations. AI has helped the development of FL, allowing secure genomic data sharing while maintaining privacy. The advancement of XAI has improved the interpretability of DL, addressing challenges related to black-box predictions [11-12].

Table 1 below highlights numerous diseases, AI and DM used, and corresponding gene/protein sequences involved in DGP.

Diseases types

Disease	Techniques Used	Gene/Protein Sequence (Full Form)
Breast Cancer	DL, Convolutional Neural Network (CNN), RF	BRCA1 (Breast Cancer Gene 1), BRCA2 (Breast Cancer Gene 2), TP53 (Tumor Protein 53)
Lung Cancer	Reinforcement Learning (RL), RF, Deep CNN	EGFR (Epidermal Growth Factor Receptor), KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog), TP53 (Tumor Protein 53)
Colorectal Cancer	k-Nearest Neighbors (KNN), SVM, Gene Ontology (GO) Analysis	APC (Adenomatous Polyposis Coli), TP53 (Tumor Protein 53), KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog)
Leukemia	SVM, Feature Engineering (FE)	BCR-ABL (Breakpoint Cluster Region-Abelson Murine Leukemia Viral Oncogene), FLT3 (FMS-like Tyrosine Kinase 3), RUNX1 (Runt-Related Transcription Factor 1)
Liver Cancer	Gradient Boosting (GB), Gene	TERT (Telomerase Reverse Transcriptase), CTNNB1

	Expression Analysis	(Catenin Beta 1), TP53 (Tumor Protein 53)
Ovarian Cancer	DL, Unsupervised Clustering	BRCA1 (Breast Cancer Gene 1), BRCA2 (Breast Cancer Gene 2), TP53 (Tumor Protein 53)
Pancreatic Cancer	Decision Trees (DT), Neural Networks (NN)	KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog), CDKN2A (Cyclin Dependent Kinase Inhibitor 2A), TP53 (Tumor Protein 53)
Alzheimer's Disease	DT, Extreme Gradient Boosting (XGBoost), Clustering	APOE (Apolipoprotein E), APP (Amyloid Precursor Protein), PSEN1 (Presenilin 1)
Parkinson's Disease	SVM, FS, Gene Network Analysis (GNA)	SNCA (Alpha-Synuclein), LRRK2 (Leucine Rich Repeat Kinase 2), PARK7 (Parkinsonism Associated Deglycase)
Huntington's Disease	XAI, TL	HTT (Huntingtin)
Amyotrophic Lateral Sclerosis (ALS)	CNN, Graph-based Learning	SOD1 (Superoxide Dismutase 1), C9orf72 (Chromosome 9 Open Reading Frame 72), TARDBP (TAR DNA Binding Protein)
Diabetes (Type 1 & Type 2)	Naïve Bayes (NB), Association Rule Mining (ARM), K-Means Clustering	INS (Insulin), GCK (Glucokinase), TCF7L2 (Transcription Factor 7 Like 2)
Cardiovascular Diseases	NN, Logistic Regression (LR)	LDLR (Low-Density Lipoprotein Receptor), MYH7 (Myosin Heavy Chain 7), PCSK9 (Proprotein Convertase Subtilisin/Kexin Type 9)
Asthma	FE, GNA	IL4 (Interleukin 4), IL13 (Interleukin 13), ADRB2 (Adrenergic Beta-2 Receptor)
Arthritis (Rheumatoid & Osteo)	Clustering, RF	PTPN22 (Protein Tyrosine Phosphatase Non-Receptor Type 22), HLA-DRB1 (Human Leukocyte Antigen DR Beta 1)
Chronic Kidney Disease (CKD)	NB, RL	APOL1 (Apolipoprotein L1), UMOD (Uromodulin), PKD1 (Polycystin 1)
Obesity	XGBoost, K-Means Clustering	FTO (Fat Mass and Obesity-Associated Gene), MC4R (Melanocortin 4 Receptor), LEP (Leptin)
Rare GD	TL, XAI	CFTR (Cystic Fibrosis Transmembrane Conductance Regulator), MECP2 (Methyl-CpG Binding Protein 2)
COVID-19 Susceptibility	NLP, RNA-seq Data, DL	ACE2 (Angiotensin-Converting Enzyme 2), TMPRSS2 (Transmembrane Serine Protease 2), IL6 (Interleukin 6)
Autoimmune Diseases (Lupus, Multiple Sclerosis)	Bayesian Networks (BN), Hybrid AI	HLA-DRB1 (Human Leukocyte Antigen DR Beta 1), PTPN22 (Protein Tyrosine Phosphatase Non-Receptor Type 22), STAT4 (Signal Transducer and Activator of Transcription 4)
Schizophrenia	NN, Principal Component Analysis (PCA), GWAS-based ML	DISC1 (Disrupted in Schizophrenia 1), NRG1 (Neuregulin 1), DTNBP1 (Dystrobrevin Binding Protein 1)
Bipolar Disorder	CNN, Deep RL	CACNA1C (Calcium Voltage-Gated Channel Subunit Alpha1C), ANK3 (Ankyrin 3), BDNF (Brain-Derived Neurotrophic Factor)
Depression	SVM, Sentiment Analysis	SLC6A4 (Serotonin Transporter), BDNF (Brain-Derived Neurotrophic Factor), FKBP5 (FK506 Binding Protein 5)

The impact of AI and DM in DGP is profound, improving early diagnosis, personalized medicine, and treatment strategies.

The challenges remain in integrating multi-omics data, ensuring model interpretability, and handling imbalanced datasets. Future research will focus on hybrid AI combining biological knowledge and computational power to refine DGP using gene sequences.

Comprehensive Analysis of Disease Prediction Using Gene Data

The literature review on DGP data explores integrating genomic data with computational models to enhance early diagnosis and risk measurement. Improvements in high-throughput sequencing, GWAS, and microarrays enabled researchers to extract informative patterns from sizeable genetic data sets. ML and DL, like SVM, CNN, and ensemble models, have pervasively been applied to DP, susceptibility inference, and investigation of GD. FS, from hybrid optimization algorithms to statistical methods, play critical roles in selecting correct biomarkers that improve model performance. Ethical concerns about genomic data privacy also mandate severe security protocols. Meta-analysis of peer-reviewed journal articles illuminates model performance, limitations, and predictions of accuracy medicine.

Comprehensive Analysis of DGP Using Gene Data

Ref.	Scope of Application	Type of Prediction	Type of Evidence	Inference	Outcome	Drawback
[13]	Melanoma skin cancer DNA damage detection	Binary classification	Genomic sequencing	The CNN-based model outperforms LR	96% accuracy in DNA damage prediction	Limited to the melanoma dataset
[14]	Degenerative disease diagnosis	Multi-modal disease classification	Imaging, genetic, and clinical	Graph-based fusion improves disease diagnosis	Outperforms state-of-the-art graph models	Computationally intensive
[15]	Alzheimer's disease classification	Disease classification	GWAS	Deep transfer learning improves SNP-based classification	89% accuracy	Dependence on GWAS dataset availability
[16]	Genomic machine learning model evaluation	Meta-analysis of model performance	Multiple genomic ML	Hyperparameter tuning and data leakage affect performance	Identifies common biases in genomic ML	Risk of overfitting in M
[17]	lncRNA-miRNA interaction prediction	Computational interaction prediction	Public databases and computational models	Review of network and sequence-based methods	Comprehensive survey with database updates	No new predictive model
[18]	Cancer classification using gene expression data	Multi-class classification	Gene expression	Fuzzy classifier improves FS	Enhanced accuracy and generalization	High-dimensional data complexity
[19]	Prostate cancer classification	Disease classification	Microarray gene expression	LSTM-DBN with optimization improves accuracy	Optimized PRC classification	Hyperparameter tuning complexity

[20]	Heart disease prediction	Multi-model classification	HD (public)	Hybrid DL outperforms traditional methods	98.86% accuracy in HD prediction	Requires large datasets for training
[21]	Microarray gene expression classification	Cancer classification	Microarray gene expression	Wilcoxon Sign Rank Sum and Grey Wolf optimized ensemble learning improve classification	100% accuracy using optimized XGBoost and CatBoost	Risk of overfitting on small datasets
[22]	Alzheimer's detection using genetic data	Binary classification (Disease/No Disease)	GWAS, SHAP explainability	SVM with SHAP enhances interpretability	89% accuracy (SVM)	Limited generalizability due to dataset bias
[23]	Leukemia prediction from gene expression	Binary classification (Leukemia/ No Leukemia)	Microarray gene expression	Hybrid ALO + PSO improves FS	87.88% accuracy (SVM)	High computational cost
[24]	Breast cancer survival prediction	Multi-class survival prediction	Multi-omic integration (gene, protein, clinical)	Genomic data enhances survival prediction	Not mentioned	Complexity in feature fusion
[25]	Diabetic Retinopathy Progression Analysis	Multi-class severity prediction	Fundus images and OCT scans	Vision transformers outperform CNNs	90.1% accuracy (ViT)	High computational demand
[26]	Gene expression classification	Multi-class classification	Gene expression (Cancer)	GRU-RNN models outperform CNN, LSTM, and hybrid models	85.7% accuracy	Limited to cancer datasets
[27]	Depression prediction	Binary classification (Depressed/ Non-Depressed)	Microarray gene expression	DP-BERT pre-trained model improves depression classification	91.2% accuracy	Generalizability concerns due to batch effects
[28]	Cancer classification	Multi-class classification	Gene expression data	DSCNN with Enhanced Chimp Optimization (ECO) improves feature selection	86.5% accuracy	Computationally intensive

[29]	Gene constraint estimation	Gene prioritization	Population genetics, ML on gene features	GeneBayes models enhance gene constraint estimation	87.9% accuracy	Limited to evolutionary analysis
[30]	Rare genetic disease diagnosis	Gene prioritization	LLM-based phenotype-gene mapping	GPT-4 attains the highest accuracy but underperforms traditional methods	88.2% AUC	LLMs require further optimization
[31]	Parkinson's disease biomarker prediction	Biomarker identification	Transcriptome and metabolic modeling	TAMBOOR improves metabolic biomarker detection	Enhanced biomarker prediction for PD	Limited to metabolic pathway analysis
[32]	Disease Symptoms prediction	Genomic Expression Classification-Based Phenotype Prediction	Topological Data Analysis	TDA-GCN-SVM model	95% Accuracy	Limited to Topological Data

3. IMPLEMENTATION OF AI FOR DGP

The genes involved in the GD of humans are identified using mutation analysis and linkage analysis, and the gene analysis is tested on the candidate gene. The data collected for biomedicine is ever-increasing, and a proficient method is needed to process it. AI, vital for DL, has dramatically succeeded in computational biology. The clinical trial, generation of a candidate gene, diagnosis, identification, and basic research are attained with the assistance of AI-based DL. Most GD are rare diseases primarily underrepresented in clinical and basic research, mainly benefiting from AI technologies.

AI is used when machines can do tasks that generally need human intelligence. It comprises ML, where machines can learn by knowledge and gain skills without the involvement of humans. DL is a subdivision of ML in which artificial neural networks (ANN) are involved in data analysis. The pattern of human behavior inspires ANN, which learns skills from massive genomic data. Gene ontology mainly focuses on the products and function of the gene, whereas terminology motivates products and genes. The unification of gene and product features across the species is attained by gene ontology. The ANN is applied from the generated gene ontology to analyze, annotate, and investigate the biomedical data.

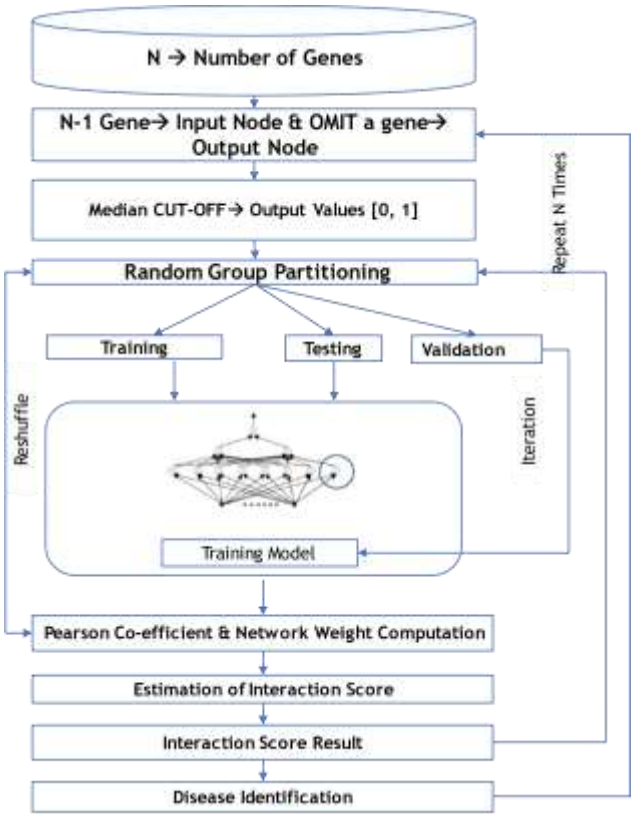
Disease caused by GD is due to the alteration in the part or whole portion of the DNA sequence. The mutation process instigates GD in one gene, mutation in several genes, factors of the environment that destroy chromosomes, and mutation in a combination of the genes. These are all the main factors that modify the gene, and it cause GD in humans. The DGP and GD from the ever-growing genomic data is a tedious method. The traditional method of DGP from the gene is complicated and time-consuming. To overcome the problems, several algorithms were developed for methods for DGP and GD-causing genes.

The GD is identified by generating ontology from the gene and applying AI-based algorithms. Currently, ontology content curation is mainly implemented in research activities related to biomedical data, and it is used in two major processes. Ontologies signify the relations and entities of diverse domains of biomedicine. The biomedical experimentalist uses ontology to interpret the data, and the data is integrated with data from other researchers, which permits the data analysis of cross-species. Both content curation and annotation are essential challenges in gene analysis. To resolve this, an AI is used.

AI is a combination of computing models, theories, and algorithms that help several things that require human intelligence, such as the perception of visuals, recognition of speech, reasoning, understanding of language, and decision-making. AI encompasses several methods: Computer Vision (CV), ML, NLP, rule-based logic, and DL. AI-based methods can speed up the analysis of massive amounts of data, leveraging patterns and giving quick results that can be used in further decision-making. Sophisticated analytical models are generated using algorithms that uncover the patterns and predict the outcomes. The arrival of big data and the increasing data demands computing power, storage, and practical data analysis. The result of the data analysis provides actionable and valuable visions.

A gene ontology is prepared with the genomic data, and AI is employed in DGP. A backpropagation NN is incorporated to design a model where the prospective collaboration among genes is achieved. The weight prediction and direction of the signal

are employed to generate a model, and its strengths are the interaction of directions and signals of the interaction link among genes. The proposed ANN is validated with Monte Carlo cross-validation to optimize generalization ability and reduce the risk of overfitting the model. The network interface and investigation of discovered genes within the scenario of the DGP highlight the extensive influence of definite genes in the genomic data with the disease. The research methodology in the literature is given below based on the diversified research. The overall mechanism of training and identifying the data from the literature is illustrated in Figure 1.



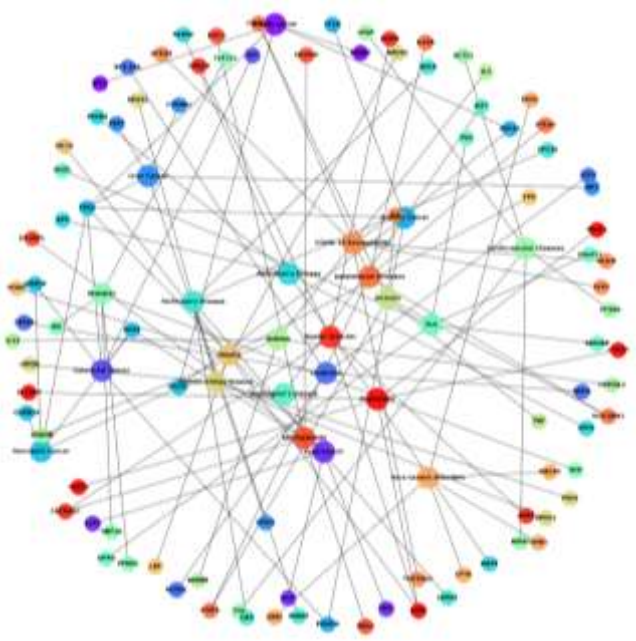
Overall Framework - AI-based Disease Prediction

4. DISCUSSION

Visualization of Similarities (VoS) Viewer 1.6.20 is a visual and analytical citation, co-citation, co-authorship, and keyword co-occurrence bibliometric network software. VoS enables the development of graphical models of the relationship between authors, papers, and keywords through clustering and network visualization. VoS is fast on standard personal computers with a multi-core processor, 16 GB RAM, and at least 100 GB storage capacity. It is compatible with MacOS 15 Sequoia. Alternatively, Python is a high-level programming language widely used for data manipulation, statistical analysis, and ML with packages like pandas, matplotlib, seaborn, and networks, further boosting its application in literature analysis. Python is adaptable to handle big data and accommodate complex algorithms, can be run on computers with a minimum of 16 GB RAM and multi-core processors, and is compatible with macOS 15

Sequoia. VoS and Python are key software in comprehensive literature analysis since they enable researchers to graphically display bibliometric data, identify trends, perform comprehensive statistical network analysis, and generate sensitive data on scientific fields and research networks.

The aim here is to visually display the correlation of genes with their diseases in a bipartite network in Figure 2. A disease is color-coded, and the relevant genes are color-coded in the same color to identify GD quickly. The network considers the GD of different diseases by visualizing common genes, such as tumor protein p53 (TP53), that are linked to different cancers. The dense layout presents a more compact visualization to make it perfect for scientists investigating GD, comorbidities, and probable therapy targets. The method improves the interpretability in genomics and bioinformatics studies.



Bipartite Gene-Disease Network: Mapping Genetic Associations Across Disorders

The reviewed studies focus on disease classification and risk prediction across several medical conditions using genomic, imaging, and clinical data. The highest reported accuracy is 100% for cancer classification using microarray gene expression data with optimized XGBoost and CatBoost, but the study raises concerns about overfitting due to small datasets. Heart disease prediction achieved 98.86% accuracy using a hybrid DL, emphasizing the advantage of multi-models but requiring big datasets. Prostate cancer classification with Deep Belief Network - Long Short-Term Memory (DBN-LSTM) optimization yielded significant improvement, though hyperparameter tuning complexity was noted. Similarly, Alzheimer's classification with Deep Transfer Learning (DTL) reached 89% accuracy, limited by GWAS dataset availability.

CNN, such as those in melanoma DNA damage detection and Parkinson's detection, reported 96% and 92.3% accuracy, respectively, outperforming traditional models. Graph-based fusion demonstrated superior performance in multi-modal disease diagnosis, though computational costs remain problematic. Vision transformers surpassed CNNs in diabetic retinopathy severity prediction with 90.1% accuracy but demanded high computational resources. Explainability-enhanced models like SVM-model-agnostic Shapley additive explanations (SHAP) for Alzheimer's detection achieved 89% accuracy but were constrained by dataset bias, which is illustrated in Figure 3.

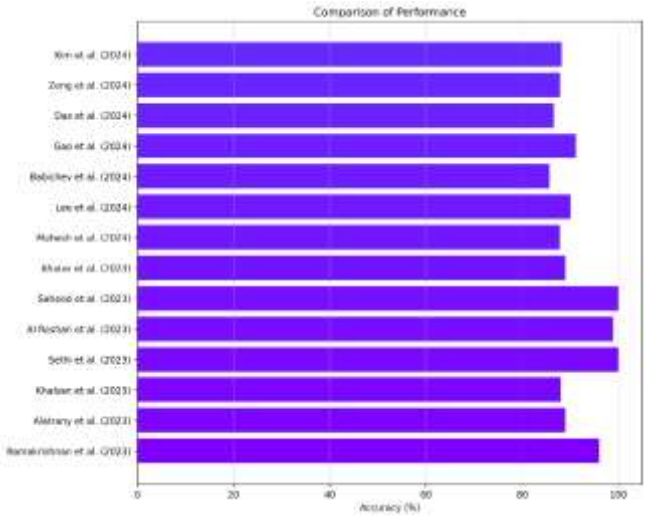


Fig. 3. Comparison of Performance

Hybrid models integrating DL and FS reliably outperformed baseline models, highlighting the importance of optimized models. However, challenges such as dataset bias, overfitting risks, and computational costs require further research to enhance

generalizability and scalability. The performance of ML, particularly classification, is primarily determined in terms of accuracy measures. The accuracy levels of several research articles reflect the performance of their respective models. The variability in the accuracies seen, with the best accuracy levels of 100% being attained by specific models, reflects the potential of well-optimized algorithms for specific tasks. However, the variability in accuracy, such as the rate of values less than 90%, reflects the inability to achieve robustness over many datasets and problem spaces. These differences can arise from the dataset quality, feature engineering, or choice of model and training parameters. However impressive, models with 100% accuracy need not always reflect the generalization potential since overfitting is possible. Accuracy must be followed by precision, recall, and F1-score to approximate such models' true performance and applicability.

5. CONCLUSION AND FUTURE WORK

The use of AI and DM for DGP has revolutionized the precision and efficacy of detecting disease-causing GD. Researchers can find complex gene-disease relationships using ML, DL, and network-based analysis tools, resulting in early diagnosis and tailor-made treatment. The challenges include integrating heterogeneous genomic data, class imbalance, and model interpretability. These challenges will be overcome through hybrid AI that combines computational capability with biological data. Integration of multi-modal data, including genomic, transcriptomic, and clinical data, will further improve the accuracy of the predictions. In addition, the creation of explainable AI and FL will provide privacy and transparency, resulting in AI-based genomic studies that are open and reliable. As the models improve, they will be a prime driver of precision medicine, giving personalized therapeutic methods to diseases.

Future research can explore hybrid AI, integrating genomic, transcriptomic, and clinical data to increase the accuracy of DGP. Emphasis on explainable AI, FL, and data bias correction will enhance the interpretability of the models and preserve privacy, thus making them clinically applicable in precision health and personalized medicine

REFERENCES

[1] Ramakrishnan, R., Mohammed, M. A., Mohammed, M. A., Mohammed, V. A., Logeshwaran, J., & Maheswaran, S. (2023, July). An innovation prediction of DNA damage of melanoma skin cancer patients using deep learning. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[2] Guo, R., Tian, X., Lin, H., McKenna, S., Li, H. D., Guo, F., & Liu, J. (2023). Graph-based fusion of imaging, genetic, and clinical data for degenerative disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[3] Barnett, E. J., Onete, D. G., Salekin, A., & Faraone, S. V. (2023). Genomic machine learning meta-regression: insights on associations of study features with reported model performance. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[4] Sheng, N., Huang, L., Gao, L., Cao, Y., Xie, X., & Wang, Y. (2023). A survey of computational methods and databases for lncRNA-miRNA interaction prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(5), 2810-2826.

[5] Khalsan, M., Mu, M., Al-Shamery, E. S., Ajit, S., Machado, L. R., & Agyeman, M. O. (2023). A novel fuzzy classifier model for cancer classification using gene expression data. IEEE Access, 11, 115161-115178.

[6] Sethi, B. K., Singh, D., Rout, S. K., & Panda, S. K. (2023). Long Short-Term Memory-Deep Belief Network based Gene Expression Data Analysis for Prostate Cancer Detection and Classification. IEEE Access.

[7] Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., & Shaikh, A. (2023). A robust heart disease prediction system using hybrid deep neural networks. IEEE Access.

[8] Sudhakar, S., Abolfazl, M., Surbhi, B., Saranya, S. S., Meshal, A., Shakila, B., Subramaniyaswamy, V., (2022), Echocardiographic Image Segmentation For Diagnosing Fetal Cardiac Rhabdomyoma During Pregnancy Using Deep Learning, IEEE Access, DOI:10.1109/ACCESS.2022.3215973.

[9] Saheed, Y. K., Balogun, B. F., Odunayo, B. J., & Abdulsalam, M. (2023). Microarray gene expression data classification via Wilcoxon sign rank sum and novel Grey Wolf optimized ensemble learning models. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(6), 3575-3587.

[10] Khater, T., Ansari, S., Alatrany, A. S., Alaskar, H., Mahmoud, S., Turky, A. & Hussain, A. (2024). Explainable Machine Learning Model for Alzheimer Detection Using Genetic Data: A Genome-Wide Association Study Approach. IEEE Access.

[11] Mahesh, T. R., Santhakumar, D., Balajee, A., Shreenidhi, H. S., Kumar, V. V., & Annand, J. R. (2024). Hybrid ant lion mutated ant colony optimizer technique with particle swarm optimization for leukemia prediction using microarray gene data. IEEE Access.

[12] Asir Chandra Shinoo, R. V., Sudhakar, S., (2024), Edge computing-based ensemble learning model for health care decision systems, Sci Rep 14, 26997. https://doi.org/10.1038/s41598-024-78225-5.

[13] Asir Chandra Shinoo R. V., Sudhakar, S., (2024), Effective clinical decision support implementation using a multi-filter and wrapper optimisation model for Internet of Things based healthcare data. Sci Rep 14, 21820. https://doi.org/10.1038/s41598-024-71726-3

[14] Munir, M. A., Shah, R. A., Ali, M., Laghari, A. A., Almadhor, A., & Gadekallu, T. R. (2024). Enhancing Gene Mutation Prediction with Sparse Regularized Autoencoders in Lung Cancer Radiomics Analysis. IEEE Access.

[15] Lee, M., Park, T., Shin, J. Y., & Park, M. (2024). A comprehensive multi-task deep learning approach for predicting metabolic syndrome with genetic, nutritional, and clinical data. Scientific Reports, 14(1), 17851.

[16] Babichev, S., Liakh, I., & Kalinina, I. (2024). Applying the deep learning techniques to solve classification tasks using gene expression data. IEEE Access.

[17] Das, A., Neelima, N., Deepa, K., & Özer, T. (2024). Gene selection-based cancer classification with adaptive optimization using deep learning architecture. IEEE Access.

[18] Eman S. S., Salah, E., Fathi E. Abd El-S., Walid, E., Nirmeen, A. E. B., Ghada, E. B., Naglaa, F. S., Sudhakar, S., Rabie, A. R., (2022), Sketch-Based Retrieval Approach Using Artificial Intelligence Algorithms for Deep Vision Feature Extraction, MDPI-Axioms, 11 (12), 663; DOI:10.3390/axioms11120663.

[19] Sudhakar, S., Abolfazl, M., Julian L., W., Ali, B., Ahlam, A., Meshal, A., Ali, A., Surbhi Bhatia, K., (2023), Improved LSTM-Based Anomaly Detection Model with Cybertwin Deep Learning to Detect Cutting-Edge Cybersecurity Attacks, Human-centric Computing and Information Sciences, 13(55). https://doi.org/10.22967/HCIS.2023.13.055

[20] Kim, J., Wang, K., Weng, C., & Liu, C. (2024). Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease. The American Journal of Human Genetics, 111(10), 2190-2202.

[21] Surbhi, B., Mohammed, A., Sudhakar, S., Pankaj, D., (2022), An efficient modular framework for automatic LIONC classification of MedIMG using unified medical language, Frontiers in Public Health, DOI:10.3389/fpubh.2022.926229.

[22] Priyadarsini, S., Carlos Andrés Tavera, R., Abolfazl, M., Vidya Sagar, P., Sudhakar, S., (2022) Automatic Liver Tumor Segmentation in CT Modalities Using MAT-ACM, Computer Systems Science and Engineering, 43(3), 1057–1068, DOI:10.32604/csse.2022.024788.

[23] Priyadarsini, S., Carlos Andrés Tavera, R., Mrunalini, M., Ganga Rama Koteswara, R., Sudhakar, S., (2022), Classification of Liver Tumors from Computed Tomography Using NRSVM, Intelligent Automation & Soft Computing, 33(3), 2022, 1517-1530, DOI:10.32604/iasc.2022.024786.

[24] Abolfazl, M., Arokia Jesu Prabhu, L., Julian, W., Dilip Kumar, S., Santhosh, J., Kousalya, K., Pallavi, S., Regin, R., Sharnil, P., Sudhakar S. (2021), Fetal health classification from cardiotocographic data using machine learning, Expert Systems, DOI:10.1111/exsy.12899.

[25] Stalin David, D., Arun Mozhi Selvi, S., Sivaprakash, S., Vishnu Raja, P., Dilip Kumar, S., Pankaj, D., Sudhakar, S., (2022), Enhanced Detection of Glaucoma on Ensemble Convolutional Neural Network for Clinical Informatics, Computers, Materials & Continua, 70 (2), 2563-2579.

[26] Razia, S., Abdul, K., Anil Gandhudi, R., Regin, R., Roy, S., Dilip Kumar, S., Mukesh Kumar, G., Hitesh, J., Ankit, K., Haritha, H., Sudhakar, S. (2021), Multilabel land cover aerial image classification using convolutional neural networks, Arabian Journal of Geosciences, 14 (1681).

[27] Muthumayil, K., Karuppathal, R., Jayasankar, T., Aruna Devi, B., Prakash, N., Sudhakar, S. (2021). A Big Data Analytical Approach for Prediction of Cancer Using Modified K-Nearest Neighbour Algorithm, Journal of Medical Imaging and Health Informatics, (8), 2120-2125 (6), DOI:10.1166/jmihi.2021.3737.

[28] Sudhakar, S., Kailash, K., Subramaniyaswamy, V., Logesh, R. (2021), Cost-effective and efficient 3D human model creation and re-identification application for human digital twins, Multimedia Tools and Applications, 10.1007/s11042-021-10842-y.

[29] Sengan, S., Priya, V., Syed Musthafa, A., Ravi, L., Palani, S., Subramaniyaswamy, V. (2020), A fuzzy-based high-resolution multi-view deep CNN for breast cancer diagnosis through SVM classifier on visual analysis, Journal of Intelligent & Fuzzy Systems, 1-14, DOI:10.3233/JIFS- 189174.

[30] Sengan, S., Arokia Jesu Prabhu, L., Ramachandran, V., Priya, V., Ravi, L., Subramaniyaswamy, V. (2020), Images super-resolution by optimal deep AlexNet architecture for medical application: A novel DOCALN, Journal of Intelligent & Fuzzy Systems, 1-14, DOI:10.3233/JIFS-189146.

[31] Shaymaa Hussein, N., Sudhakar, S., Joel Sunny, D., Serwes, B., Saravanan, V., Veeramallu, B. (2025), The Diagnosis of Heart Attacks: Ensemble Models of Data and Accurate Risk Factor Analysis Based on Machine Learning, Journal of Machine and Computing, 5(1), 589-599.

[32] Narender M, Karrar S. Mohsin, Ragunthar T, Anusha Papasani, Firas Tayseer Ayasrah and Anjaneyulu Naik R (2024), Machine Learning for Genomic Expression Classification-Based Phenotype Prediction in Topological Data Analysis, Journal of Machine and Computing 4(4) (2024)

..