

ORIGINAL ARTICLE

A Novel GNN-Bi-LSTM Framework with Attention Mechanism for Detecting Rare Genetic Disorders

Osama R. Shahin ¹, Raed N. Alabdali ¹, Mohamed N. Ibrahim ², Wael Alzahrani ²,
Eman Fawzy El Azab ², Eman Elsayed Bauomey Mohamed ^{3,4}

¹ Department of Computer Science, College of Computer and Information Sciences, Jouf University, Al Jouf, Saudi Arabia

² Department of Clinical Laboratory Science, College of Applied Medical Sciences - Qurayyat, Jouf University, Al Jouf, Saudi Arabia

³ Nursing Department, College of Applied Medical Sciences - Qurayyat, Jouf University, Al Jouf, Saudi Arabia

⁴ Nursing Department, Faculty of Nursing, Ain Shams University, Cairo, Egypt

ABSTRACT

Background: The growing population in the world and inadequate health care guidelines have caused the onset of sporadic genetic disorders that need effective diagnostic and prognostic methods. However, the existing systems cannot handle the multisystem nature of genetic information resulting in delays of formal identification and therapeutic procedures.

Methods: The proposed study aims at proposing a more reliable model, GNN-Bi-LSTM-AM, that incorporates Graph Neural Networks (GNN), Bi-LSTM, and attention module to improve the process of identifying and diagnosing rare genetic disorders (RGDs). In this model, GNN is used to describe the genetic factors' associations with other factors, whereas Bi-LSTM models the health data over time. The attention mechanism puts emphasis on significant genetic markers, which improves predictive ability and minimizes interpretation challenges. The framework makes use of complex data such as maternal health history and parental genetic inheritance. In particular, the GNN learns the association between patients on the basis of genetic similarities, and Bi-LSTM computes temporal details of health information, and the attention module is applied to pay attention to important elements.

Results: The results of the experiment show that the proposed GNN-Bi-LSTM-AM model predicts the accuracy of 98.5, precision of 96, recall of 96, and an F1-score of 96, which is much better than traditional machine learning methods. The framework has also good efficiency in that the processing time was 7.85 seconds.

Conclusions: The GNN-Bi-LSTM-AM framework suggested is a great step towards predictive analytics of rare genetic diseases. The model can be used to provide promising applications in early detection and clinical diagnosis by combining genetic relationships, sequential health data, and feature relevancy. Therefore, it is practical in enhancing treatment interventions and patient outcomes in genetic medicine.

(Clin. Lab. 2026;72:xx-xx. DOI: 10.7754/Clin.Lab.2025.250934)

Correspondence:

Mohamed Nabil Ibrahim, Ph.D.
Clinical Laboratory Science Department
College of Applied Medical Sciences, Jouf University
Qurayyat
Saudi Arabia
Email: Mnmicro2010@gmail.com

KEYWORDS

rare genetic disorders, Graph Neural Network, Bi-LSTM, attention mechanism, predictive analytics, early detection

INTRODUCTION

Rare diseases (RDs) are defined as conditions with an incidence of fewer than 5 per 10,000 individuals, with approximately 80% being genetically determined [1]. Their diagnosis poses significant challenges due to

broad phenotypic variability. The same phenotype can result from distinct genetic etiologies, while dissimilar phenotypes may involve overlapping gene networks [2,3]. Consequently, reliance on clinical manifestations alone often limits accurate genetic analysis.

Congenital disorders, encompassing autosomal dominant, recessive, chromosomal, and polygenic abnormalities, result from alterations in DNA that manifest through diverse physical and psychological symptoms [4,5]. These disorders impose a heavy burden on healthcare systems because of their diagnostic complexity and the lack of cost-effective, scalable tools for personalized genetic evaluation. The World Health Organization estimates that over 7,000 distinct rare genetic disorders affect hundreds of thousands of children globally [6]. Timely diagnosis enables early therapeutic intervention, potentially reducing complications, improving life expectancy, and enhancing overall quality of life. Early identification also facilitates the detection of genetic carriers, allowing informed family planning and the prevention of hereditary transmission - an essential public health objective [7,8].

Despite significant progress in computational genomics, existing models for predicting genetic disorders remain limited. The primary issue lies in the opacity of most machine learning models, which lack interpretability for clinicians and policymakers [9]. These models often provide single-point predictions without explaining the underlying biological rationale, hindering clinical decision-making. Moreover, scalability remains a challenge as genomic datasets expand, leading to issues such as overfitting, high computational complexity, and diminished generalizability [10]. Traditional algorithms frequently overlook the structural and relational information inherent in genetic datasets, thereby underutilizing critical genomic interconnections that could enhance predictive accuracy [11].

To address these limitations, this study introduces a novel hybrid computational framework that integrates Graph Neural Networks (GNNs) with Bidirectional Long Short-Term Memory (Bi-LSTM) networks and an attention mechanism. The proposed GNN-Bi-LSTM model is designed to capture both spatial genetic relationships and temporal dependencies in genomic and clinical data. The GNN component maps genetic and phenotypic interactions across patients, identifying latent connections that conventional models fail to detect. Simultaneously, the Bi-LSTM architecture captures sequential dependencies in longitudinal health data, while the attention module prioritizes the most relevant genetic and clinical features, enhancing model transparency and interpretability.

This hybrid approach achieves an accuracy rate of 98.5% in predicting rare genetic disorders, outperforming conventional machine learning methods. Beyond accuracy, the model's interpretive structure facilitates clinical translation by providing insight into the gene-phenotype interplay. Integrating genetic, environmental, and clinical data establishes a foundation for individual-

ized medicine and precision health strategies. Early detection and prediction of genetic disorders not only mitigate morbidity and mortality but also optimize resource allocation and reduce lifetime treatment costs.

Furthermore, the framework supports the field of pharmacogenomics by enabling the disaggregation of genetic data based on patient history and familial inheritance patterns. This enhances personalized therapeutic planning and the development of customized treatment pathways. By bridging the gap between complex genomic data and actionable clinical insights, the proposed system contributes to advancing predictive healthcare and supporting global efforts toward personalized, preventive, and participatory medicine.

In addition to its theoretical contributions, the study offers practical implications for integrating AI-driven diagnostic models into real-world clinical systems. The scalability and interpretability of the proposed architecture make it a potential foundation for future research on gene-disease networks, contributing to global scientific databases and personalized healthcare frameworks. In summary, the paper presents:

- A new multi-faceted deep learning model combining GNNs, Bi-LSTMs, and attention mechanisms for rare genetic disease prediction.
- Enhanced interpretability through modeling of gene relationships and feature prioritization, enabling clinician trust and transparent diagnostics.
- Integration of genetic, clinical, and environmental data to support personalized medicine and early disease management.
- Demonstrated high predictive accuracy (98.5%), emphasizing its superiority over traditional approaches.
- Implications for early diagnosis, prevention, and genetic counseling, ultimately improving population health outcomes.

The remainder of the study is structured as follows: Section 2 reviews previous methodological limitations; Section 3 defines the research gap; Section 4 presents the proposed framework; Section 5 details experimental findings; and Section 6 concludes with key insights and directions for future research.

MATERIALS AND METHODS

Deoxyribonucleic acid (DNA) is a complex, thread-like molecule located in the nucleus, encompassing approximately 20,000 - 30,000 genes that encode the hereditary information of humans. Even minor mutations in DNA can alter normal cellular functions, leading to abnormal gene expression and the manifestation of genetic disorders. These disorders are broadly categorized as chromosomal, combined genetic, and single-gene or monogenic diseases. Given their complexity, accurate identification and classification of such disorders require more advanced and reliable computational methods.

To enhance diagnostic accuracy and performance, Nandhini and Tamilpavai [12] proposed the Elephant Herd Optimization-Whale Optimization Algorithm (EHO-WOA), designed to optimize the Stacked ResNet-BiLSTM model. This hybrid algorithm integrates evolutionary and metaheuristic principles to refine feature selection and parameter tuning. Applied to gene expression and genotype data, the enhanced model successfully detected rare genetic syndromes such as Prader-Willi, Rett, and Angelman. The model also demonstrated robustness in identifying DNA-related abnormalities including Marfan syndrome and early-onset morbid obesity. Nonetheless, one limitation is the model's potential to overfit small but feature-rich datasets. Moreover, the symbolic modeling proposed may restrict generalization to disorders excluded during training, emphasizing the need for scalable validation strategies.

Another significant contribution in genetic prediction relates to polygenic scores (PGSs) derived from genome-wide association studies (GWASs). Blood cell traits - markers of immune and systemic health - have been associated with numerous genetic loci, supporting a polygenic model of inheritance. However, the construction of clinically reliable PGSs remains a methodological challenge. Xu et al. [13] compared six approaches for developing PGSs across 26 blood cell traits, including traditional pruning and thresholding alongside five machine learning methods: LDpred2, Elastic Net (EN), Bayesian Ridge (BR), Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNNs). Utilizing data from the UK Biobank and INTERVAL cohorts, EN and BR achieved superior predictive accuracy, outperforming other methods by 10 - 23% in external validations. The findings also highlighted sex-specific genetic interactions, revealing that certain PGSs displayed significant correlations with age-related and sex-dependent variations in blood traits. These results underscore the potential of machine-learning-optimized PGSs in disease risk stratification; however, the study's reliance on only two deep learning models limits generalization across diverse neural architectures.

A major challenge in biomedical research lies in accurately predicting genome disorders responsible for multivariate diseases such as cancer, dementia, cystic fibrosis, diabetes, and Leigh syndrome, all of which contribute to substantial global morbidity and mortality. Traditional predictive frameworks, predominantly theoretical or descriptive, have evolved with the availability of large-scale genomic datasets and computational resources. Prior research employing supervised and unsupervised algorithms often focused on binary classification of genetic sequences, neglecting patient history and longitudinal health data, leading to limited accuracy. To address these deficiencies, Atta-Ur-Rahman et al. [14] developed the Advanced Genome Disorder Prediction Model (AGDPM), structured upon AlexNet architecture. This model effectively classified multiple categories of genetic diseases including single-gene inheri-

tance, mitochondrial inheritance, and multifactorial conditions - with moderate-to-high performance across key statistical metrics. Despite its promise, the study's dataset was relatively small and limited to specific disorders, constraining its external validity. Future expansions incorporating broader genomic categories and additional deep learning techniques could enhance predictive robustness.

Genetic diseases, particularly mitochondrial and multifactorial disorders, affect millions globally and often arise from complex interactions involving inherited mutations, environmental stressors, or physiological deficiencies such as anemia and low white blood cell counts. Early detection is essential, especially among young adults, as these diseases can severely compromise quality of life. Integrating advances in computational biology, Rahman et al. [15] introduced an Internet of Medical Things (IoMT)-based framework employing Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms for genetic disease prediction. Results revealed that the SVM algorithm achieved an overall accuracy of 94.99% and testing accuracy of 86.6%, outperforming existing methods. Nonetheless, the study lacked a multiclass classification approach encompassing major disorders like cancer, dementia, and diabetes, which could enhance its clinical relevance. Future improvements should focus on expanding the dataset and refining multiclass predictions to enhance applicability in precision medicine.

Although GWASs have proven useful in identifying genetic variants linked to disease susceptibility, they often fail to directly associate variants with their regulatory mechanisms. To address this limitation, Arloth et al. [16] proposed DeepWAS, a deep learning framework that extends traditional GWAS by incorporating regulatory effect predictions for each variant. The model links single nucleotide polymorphisms (SNPs) to cell-type-specific chromatin features, allowing the identification of regulatory SNPs (dSNPs) involved in complex disorders such as multiple sclerosis (MS), major depressive disorder, and even height-associated traits. DeepWAS outperformed conventional GWAS models, achieving higher prediction accuracy for 91% of MS-specific dSNPs and establishing significant associations with quantitative trait loci (eQTLs) and methylation QTLs. Despite its high interpretability, the model's limitation lies in its dependence on chromatin data available for only a few cell types, restricting its generalizability to other tissues.

Furthermore, the identification of essential genes - those critical for organismal survival - has become a key computational objective. While experimental RNA interference and knockout screenings remain the gold standard, they are resource-intensive and time-consuming. Aromolaran et al. [17] developed a machine learning framework to predict essential genes in *Drosophila melanogaster*, utilizing an extensive feature set encompassing 27,340 attributes, including nucleotide and protein sequences, gene sets, protein-protein interactions, conser-

vation indices, and functional annotations. Cross-validation demonstrated a notably higher accuracy than protein sequence-only models. Feature analysis indicated that network topology, functional annotations, and sequential characteristics contributed most significantly to predictive success. The framework was further validated on human datasets, confirming its translational potential. However, its dependency on species-specific datasets limits extrapolation to other organisms, indicating the need for generalized, multi-species gene essentiality models.

Collectively, the reviewed literature demonstrates significant progress in leveraging artificial intelligence for genetic disorder identification and gene essentiality prediction. Models such as EHO-WOA ResNet-BiLSTM [12] and AlexNet-based AGDPM [14] exhibit superior accuracy in detecting rare and multifactorial disorders. Machine-learning-optimized PGSs [13] advance the prediction of blood cell traits and disease risk, while DeepWAS [16] bridges the gap between genomic variation and regulatory function. Similarly, integrative frameworks such as IoMT-SVM [15] and feature-rich gene essentiality models [17] extend predictive power across diverse biomedical domains.

Nevertheless, several limitations persist. Overfitting remains a key concern in small genomic datasets; model interpretability and scalability are often constrained; and the majority of existing frameworks remain specialized for specific disease subsets. Moreover, limited generalization across species and cell types underscores the necessity of cross-domain integration and multiclass classification in future research. The collective evidence underscores the growing potential of hybrid AI architectures in genomic prediction, yet it equally highlights the critical need for explainable, scalable, and transferable computational models capable of supporting precision medicine and genomic healthcare at population scale.

Research gap

The absence of adequate research in the field of genetic disorder prediction demonstrates a number of essential issues that the study in question is aimed at. Existing methods of solutions are supplemented by low predictive accuracy, preconditioned by the character of genetic information and interaction along with the cases of false diagnosis and inadequate treatment of the patients. Also, some of the traditional algorithms cannot be interpreted in any way, and a clinician needs to have a complete understanding of the reasoning behind any and all the predictions provided, which limits the variety of areas where it can be used. The other weakness is on their methods of addressing time series data; the existing methods are not usually able to utilize timely health data, which is relevant to the monitoring of developmental genetic diseases [18]. Moreover, genetic data are not integrated with clinical and environmental parameters to a sufficient extent, thus limiting the possibilities of both making more comprehensive forecasting and understanding the inheritance process. Substantial capa-

bilities of graph neural networks (GNNs) for predicting complex genotypic interactions have not been fully explored while analyzing the effects of specific rare genetic disorders. Missing values represent considerable difficulty when analyzing genetic datasets, as there is no adequate solution for inducing these gaps. Filling these gaps in an effort to have improved accuracy in the prediction of rare GDs and better interpretation and incorporation of multiple heterogeneous data types in the modeling process, this study proposes the GNN-Bi-LSTM model with attention.

Research mechanism

The approach to identifying and forecasting the development of multidisciplinary pathological conditions based on the presence of rare genetic diseases uses a combination of GNN and Bi-LSTM networks with attention to improving feature relevance. Starting with the processing of the genetic and health data, normal data pre-processing stages such as handling missing values imputation are applied, then the data is cleaned and normalized, and then an initial graphical representation is developed such that each patient is a node and connections between nodes are based on genetic similarity or family ties. The given aspect of the GNN component demonstrates a satisfactory representation of the interactions with which patients present themselves, therefore, allowing the inheritance of the genetic pattern to be developed. Subsequently, the obtained sequence of health data represented by heart rate and respiratory rate as the constructed graph is sent through the Bi-LSTM layer. The layer assists in capturing temporal dependencies needed to learn how changes in the set of health indicators take place over time. This is followed by the incorporation of the attention mechanism whereby focus is scaled and centered on the most significant aspects so that when learning, the major genetic markers and history of the patients are accorded the central position. By integrating all these elements into the integrated architecture, the model can enjoy the benefits of the GNNs and Bi-LSTMs to analyze genetic data. The model is trained in an appropriate manner of optimization that would minimize the implementation errors. The results of the given model are compared to the traditional machine learning algorithms using performance indicators. The specified enhanced methodology does not only enhance the predictive potential of individual genetic disorders but also customizes the input, thus, launching new opportunities in the field of clinical diagnostics and creation of new early intervention strategies.

Data collection

To learn the patterns and the rate of genetic disorders, it is possible to use free datasets that are found on the Kaggle site. These datasets tend to be abundant in the health and genetic history of the patient and could include people with numerous illnesses such as inherited illnesses. It might include such spheres as genetic vulnerability, family predisposition to a disease, and envi-

ronmental circumstances that might influence the process of the disease. These data points are very important in studying genetic transmission and other environmental and health facets that define the onset and the course of genetic diseases. By such databases, one is able to build prediction models to estimate the susceptibility of a population to certain genetic disorders for purposes of diagnosing them early so that poor outcomes may be avoided. In the following section, the author provides an overview of the main data fields usually linked in these Kaggle open-source databases so severe genetic disorders may be identified or predicted [19].

Data pre-processing

While addressing the problem of the identification and prognosis of low-incidence hereditary diseases when using GNN-Bi-LSTM modified by an attention mechanism, pre-processing the datasets is a critical step. The data used usually comes from genetics and medical backgrounds, which often have noise and incomplete records and contain high-dimensional characteristics. To achieve this, genetic and health data needs to undergo a pre-processing step in which the input data is prepared in an appropriate manner for model training and increases the accuracy of resulting predictions. The two techniques that stand out in this process are Min-Max normalization and missing value handling. To build an accurate and robust predictive model, the following pre-processing steps are essential:

Min-max normalization

Min-max normalization is a scaling technique used to transform features to a common scale, typically between 0 and 1. This is particularly important in genetic datasets where features may have different units and ranges, such as heart rate, respiratory rate, and various genetic markers. The formula for Min-max normalization is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

Since all the input features' values are scaled between 0 and 1 by the Min-max normalization, the model converges faster at the time of training. The largest benefit comes to algorithms that are dependent on the scale of input data, because, for example, indices like respiratory rates or genetic differences must be equally significant for the machine learning process.

Handling missing value imputation

Data deletion is common in genetic databases due to entry errors, incomplete records, or limited test Capacity, necessitating effective handling of missing values to improve predictive accuracy. Imputation techniques such as mean or median imputation address gaps in continuous variables (e.g., heart or respiratory rates), while mode imputation suits categorical data like inherited conditions. However, excessive data removal can erase valuable patterns and bias outcomes. Proper preprocess-

ing - data cleaning, imputation, and normalization - optimizes training of the GNN-Bi-LSTM model, enhancing its ability to identify rare genetic disorders accurately while ensuring that the extracted insights remain scientifically valid and representative of underlying genetic relationships.

GNN for genetic relationship extraction

Graph Neural Networks (GNNs) effectively model relationships between patients and genes, making them ideal for structured genetic data analysis [20]. In this framework, patients are represented as nodes and genetic similarities as edges, enabling the GNN to capture inheritance patterns and relational dependencies crucial for accurately diagnosing and predicting genetic disorders.

Graph representation

In the GNN framework, each node represents a patient characterized by genetic, physiological, and demographic attributes (e.g., mutations, heart rate, age). Edges denote genetic or familial relationships, capturing shared mutations, inheritance patterns, or environmental influences. These graph-based connections enable the model to learn complex dependencies between patients, thereby improving the accuracy of rare genetic disorder prediction and classification.

1. Input features: Each patient's features are first embedded in the form of genetic mutations, parental inheritance, or clinical measurements. The characteristics are needed to represent every patient in the graph.
2. Neighborhood aggregation: GNNs successively aggregate the information of the neighbors of a node (i.e., other patients that an edge connects it to). This will enable the representation of each patient to be informed not only based on their own data but also based on their genetic links with other patients. As an illustration, two patients that share gene mutation that is associated with a particular disease will be considered by the GNN, and the connection will be used to improve the representation of nodes.
3. Message passing: In every iteration, the GNN messages between interconnected nodes. This enables the transmission of genetic information through the network which makes each genetic profile of a patient updated according to his/her relationship with other patients in the dataset.
4. Feature extraction: Once neighborhood aggregation and message passing have occurred several times, the representation generated by each node will reflect the personal genetic profile of the patient as well as the impact of their ties on other patients. This enables one to extract sophisticated patterns that are critical in determining genetic disorders.
5. Global representation: The final node embeddings, which now incorporate genetic information from both the individual patient and their genetic relation-

ships, can be used as input to downstream tasks, such as predicting the likelihood of a genetic disorder.

Graph Neural Networks (GNNs) present remarkable advantages in detecting and predicting rare genetic disorders by efficiently modeling complex relationships among patients through shared genetic and clinical information. They capture dynamic inheritance mechanisms and intricate gene-gene and gene-environment interactions, allowing information to flow between genetically similar patients (nodes). This capacity enables GNNs to uncover hidden inheritance structures, recognize unique genetic signatures, and deliver personalized predictive insights. Moreover, GNNs provide individual recommendations by representing patients according to genetic similarity and mutation patterns, which significantly improves diagnostic precision compared to conventional analytical methods. Their ability to infer multifactorial dependencies enhances the understanding of heritability and disease transmission patterns. When combined with Bidirectional Long Short-Term Memory (Bi-LSTM) networks and attention mechanisms, GNNs further minimize prediction errors, enhance interpretability, and generate robust, clinically meaningful outcomes. This hybrid approach establishes a powerful framework for accurate, explainable, and personalized modeling of complex rare genetic disorders.

Bi-LSTM Layer for detection of genetic disorders

The Bi-LSTM layer, a bidirectional long short term memory layer assists in the detection and prediction of these rare genetic disorders, particularly when it is applied to sequential health data which is time series data of heart rate, respiratory rate and other such measures. In comparison to LSTM which supports data purely in only one direction in a one-way traffic system, Bi-LSTM supports the input in both directions thus becoming sensitive to temporal relationships of the past and future. This is a two-way bidirectional approach, which implies that Bi-LSTM proves highly useful in health data analysis when both the past and the future have an impact on the health of a patient to determine it more accurately. Calculation of long term and rules for short-term memory gates are mentioned in eqn. (2), (3), (4), (5), (6), and (7) aiding the calculation of the hidden state h_t for forward and reverse state.

$$f_t = \sigma(W_f \cdot [h_t - 1, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i) \quad (3)$$

$$C'_t = \tanh(W_c \cdot [h_t - 1, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * C'_t \quad (5)$$

$$O_t = \sigma(W_o \cdot [h_t - 1, x_t] + b_o) \quad (6)$$

$$h_t = O_t * \tanh(C_t) \quad (7)$$

The Bi-LSTM layers are of utmost importance during prediction phase where future health status is predicted based on current and past health sequence. Past and future added health information enables the model to

make more precise probability estimates on the development of genetic diseases in a certain patient. This renders Bi-LSTM quite efficient in tracing the progress of a disorder over time and earlier identification of symptoms that would help to add to the interventions and tailor-made therapies. Bi-directional LSTM - it is particularly applied to sequential data in the predictors and as we know the course of health can be considered a process that is constantly undergoing change, with this layer making one significant component of the models developed to identify exogenous conditions and make predictions of rare genetic disorders. Combined with other factors, such as Graph Neural Networks (GNNs), it complements the model of genetic features learning and health indicators analysis powerfully, being more accurate and timelier.

Integration of Attention Mechanism for Enhancing Genetic Disorder Detection in GNN-Bi-LSTM Models

The proposed architecture presents a robust and Intelligent framework for predicting rare genetic disorders, integrating the strengths of Graph Neural Networks (GNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. This hybrid model allows both architectures to complement one another - GNNs excel in learning relational and structural dependencies within genetic data, while Bi-LSTMs are effective in modeling temporal and sequential variations in physiological parameters. Together, they enhance the interpretability, reliability, and accuracy of predictive outcomes.

In the initial stage, the GNN processes genetic data represented as a graph, where each node denotes a patient characterized by genetic, clinical, and demographic features, and edges represent genetic or familial similarities, such as shared mutations or inheritance patterns. Through iterative information propagation across nodes, the GNN identifies inter-patient dependencies and complex inheritance trends that are often missed by conventional learning methods. This step enables the model to understand both individual genomic signatures and population-level relational dynamics.

Subsequently, the extracted node embeddings are fed into the Bi-LSTM layer, which specializes in analyzing sequential health data such as heart rate and respiratory rate. The bidirectional configuration allows the model to capture contextual dependencies across both past and future steps, reflecting how physiological signals evolve in connection with genetic predispositions.

The attention layer in the proposed framework always gives more weight to the biologically significant variables, which are maternal illness history, parental inheritance patterns, and the existence of birth defects. Such a weighting manner is indicative of the fact that the underlying biological relevance of these features are well represented in the prediction of rare genetic disorders by the model. The attention mechanism shows that the assignment of computational feature precedence to clinically significant variables is consistent with known ge-

Table 1. Dataset description.

Data field	Description
Gene inherited from mother's side	Indicates whether the genetic disorder is inherited from the mother's side.
Gene inherited from father's side	Indicates whether the genetic disorder is inherited from the father's side.
Paternal gene	Reflects any gene defects inherited from the paternal side of the family.
Respiratory rate (breaths/min)	A patient's breathing rate is an important health indicator in genetic disorder prognosis.
Heart rate (beats/min)	The patient's heart rate, which can indicate cardiovascular involvement in some genetic disorders.
Parental consent	Indicates whether the patient's parents have consented to genetic testing or treatment plans.
Follow-up risk level	Represents the level of risk based on the patient's health and progression of the disorder.
Folic acid details (peri-conceptional)	The patient's intake of folic acid during pregnancy is important for preventing neural tube defects.
History of serious maternal illness	Reflects significant health conditions in the patient's mother that could affect genetic risk.
History of radiation exposure (X-ray)	Any prior radiation exposure experienced by the patient, potentially increasing genetic mutation risks.
History of substance abuse	Indicates whether either parent has a history of substance abuse, which may contribute to genetic risks.
Assisted conception details	Records the type of infertility treatments used (e.g., IVF), which can sometimes impact genetic outcomes.
Birth defects	Identifies any birth defects observed in the patient, relevant for understanding inherited conditions.

Table 2. Sources of genetic defects in patients.

Source of genetic defect	Percentage (%)
Inherited from father	40.0
Inherited from mother	35.0
Paternal family gene	25.0

Table 3. Relationship between birth defects and parental consent.

Parental consent	Birth defects present	Birth defects absent	Total cases
Yes	150	650	800
No	200	100	300

Table 4. Effect of maternal health on likelihood of genetic disorders in offspring.

Maternal health history	Patients with genetic disorders	Patients without genetic disorders
Serious illness present	700	200
No serious illness	150	750

Table 5. Time cost analysis for the different machine learning algorithm.

Machine learning algorithm	Time cost (s)
ANN	75.34
AdaBoost	25.27
SVM	1,235.48
KNN	24.71
Proposed GNN-Bi-LSTM-AM	7.85

Table 6. Performance of proposed GNN-Bi-LSTM-AM model.

Metric	Value (%)
Accuracy	98.5
Precision	96.0
Recall	96.0
F1-Score	97.0

Table 7. Performance evaluation of the proposed approach.

Methods	Accuracy	Precision	Recall	F1-Score
SVM	90	89	90	89.5
Naive bayes	95	93	92	93
K-nearest neighbor	94	92	91	92.5
Proposed (GNN-Bi-LSTM-AM)	98.5	96	96	97

netic and physiological predictors, thus providing more interpretability as well as strengthening the biological plausibility of the predictive results.

To enhance interpretability and precision, an attention mechanism is incorporated. This mechanism dynamically assigns weights to critical genetic and clinical features, highlighting the most influential patterns that contribute to disease prediction. It allows clinicians to visualize which variables play the most significant role in decision-making. Furthermore, pooling layers are utilized to consolidate features, reduce dimensionality, and strengthen the representation of essential information. By merging the relational insights of GNNs with the temporal modeling capabilities of Bi-LSTMs, and refining both through attention and pooling, this architecture delivers a comprehensive, transparent, and high-performing solution for early diagnosis and personalized prediction of rare genetic disorders.

RESULTS AND DISCUSSION

This section contains the analysis of the recently developed GNN-Bi-LSTM model with attention mechanism to reveal and predict the presence of rare genetic diseases. The work of the experiments was carried out by Python. The model was derived and tested on a large genome database that included sequential health records and genetic patient correlations. The findings are measured in several objective indices starting with accuracy and followed by precision, recall, F1-score, as well as the area under the ROC curve. A comparison with other models is made to show the enhancements introduced by the combination of the attention mechanism with the GNN and Bi-LSTM structure.

Prevalence of Genetic Defects by Parental Inheritance

Concerning the information about genetic defects according to the type of inheritance, the circulation of the genetic disorders inherited from the mother, father as

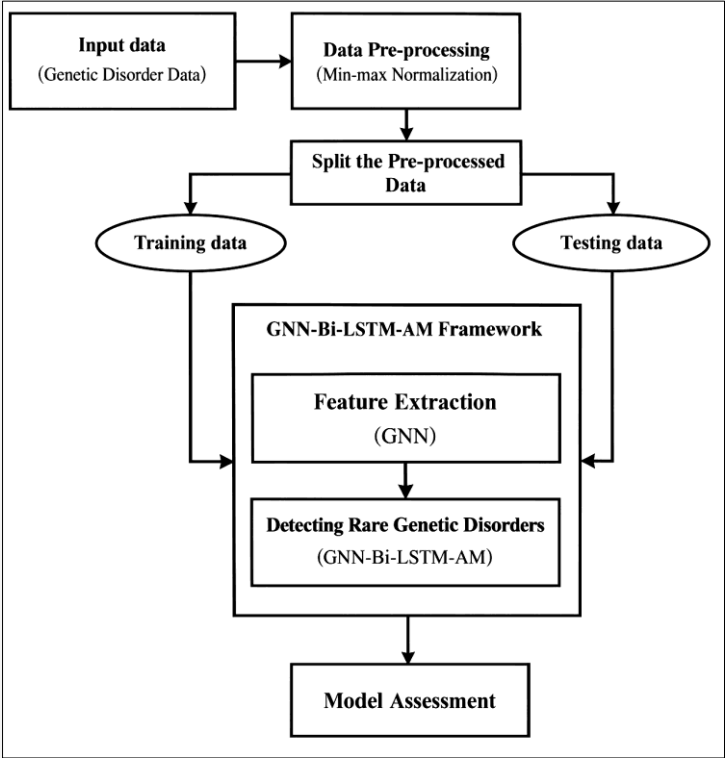


Figure 1. Workflow of the proposed framework.

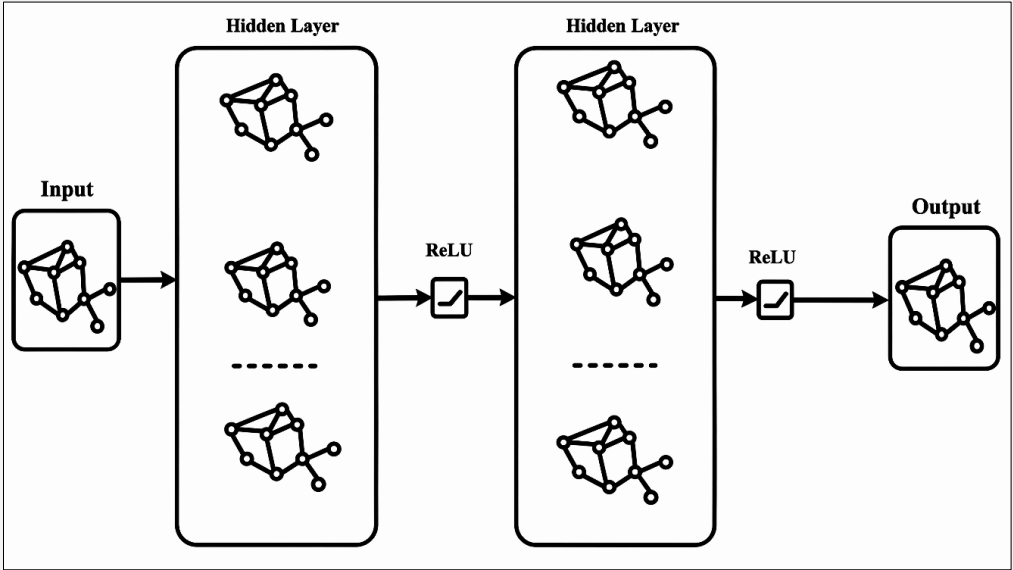


Figure 2. Architecture of GNN.

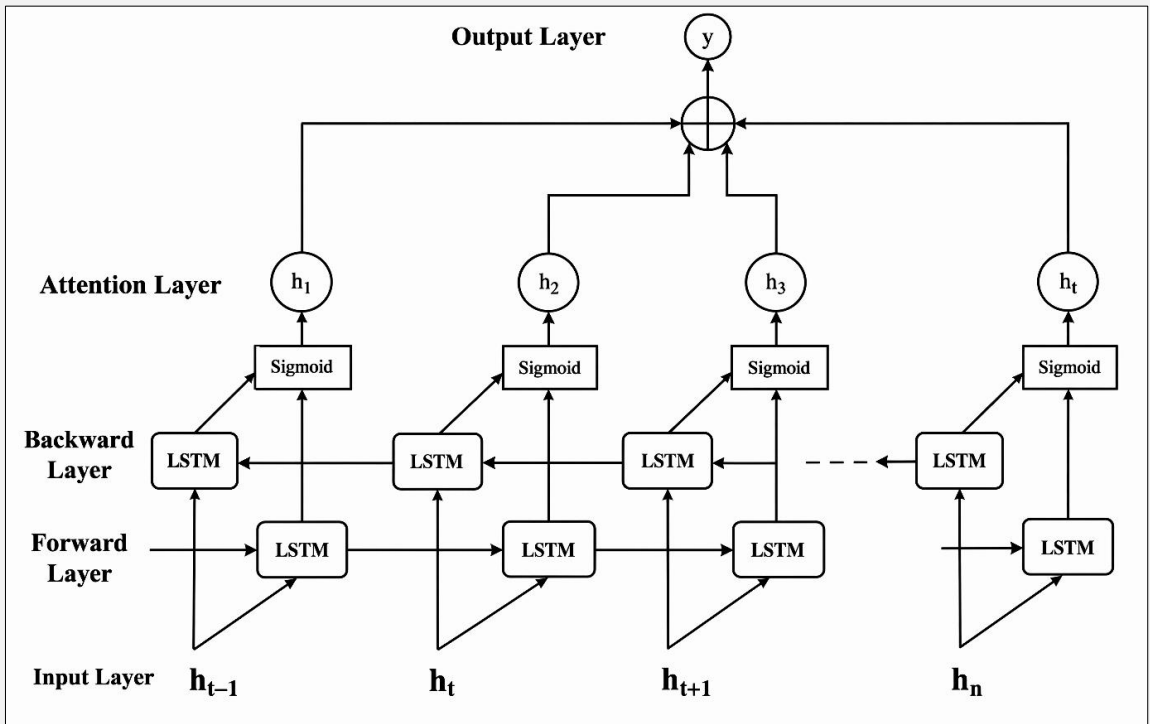


Figure 3. Architecture of Bi-LSTM-AM.

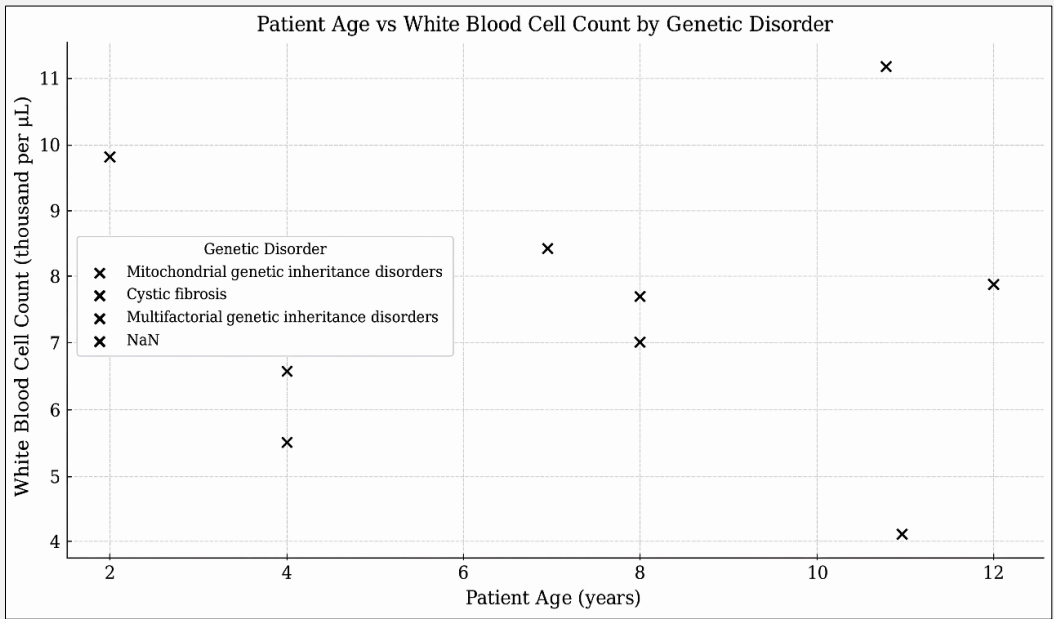


Figure 4. Analysis of genetic disorder.

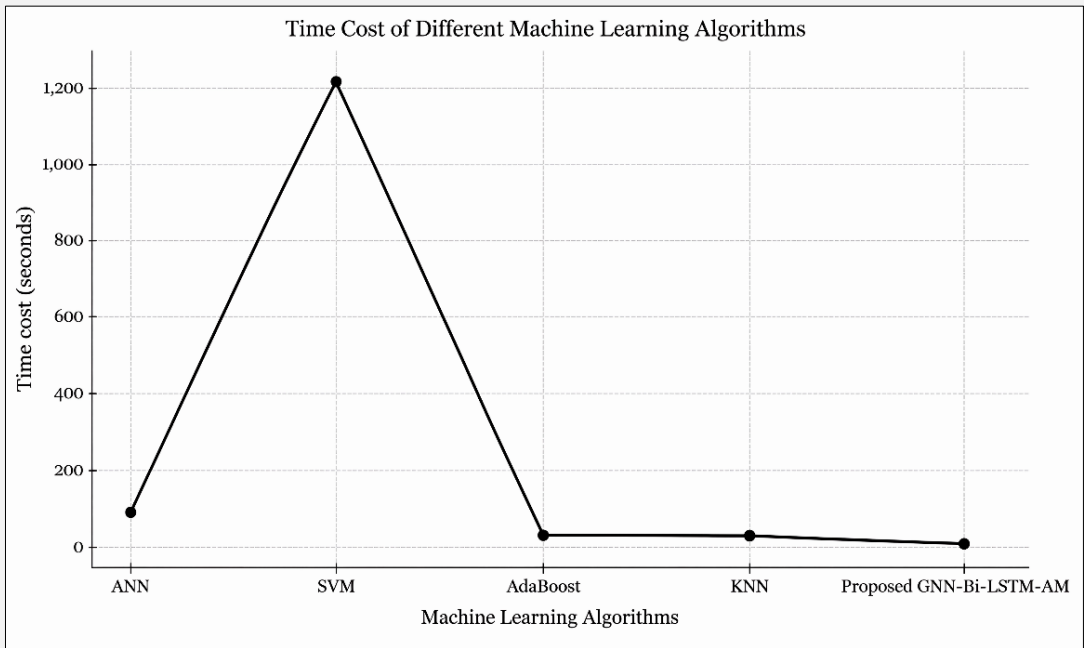


Figure 5. Cost effectiveness assessment.

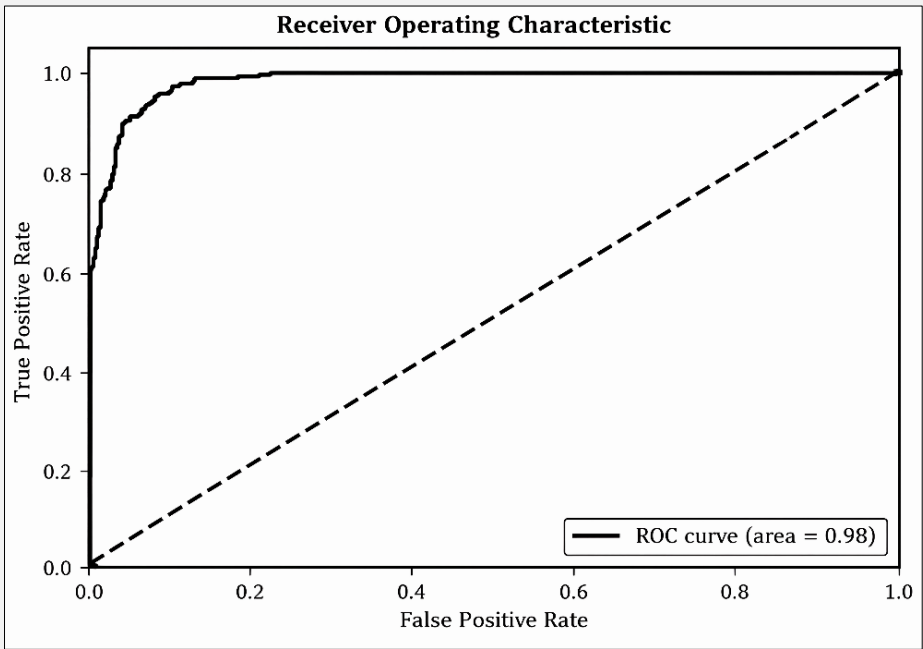


Figure 6. ROC curve.

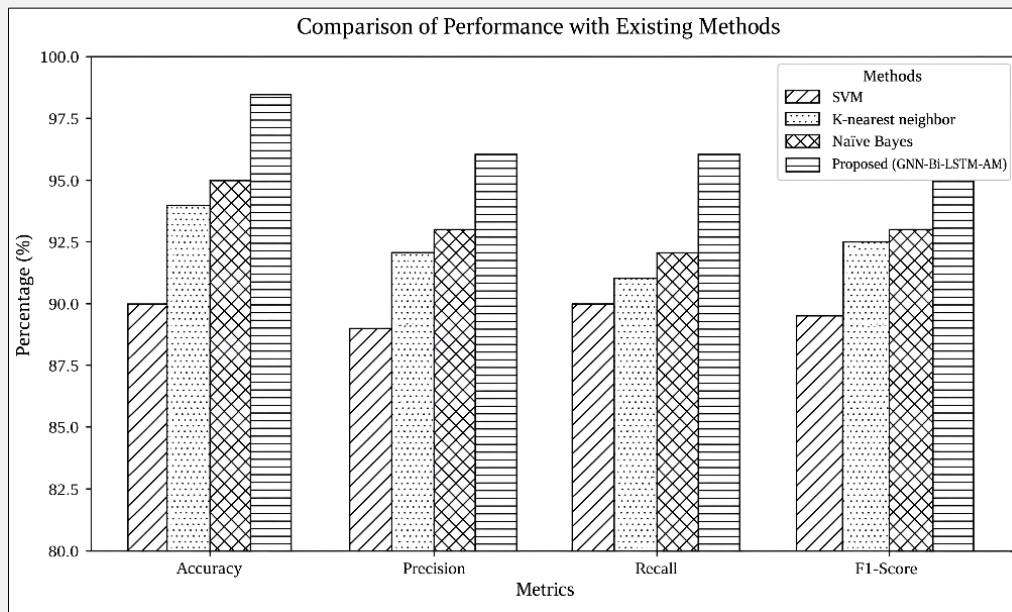


Figure 7. Comparative assessment.

well as a paternal line is visualized in Figure 4. Among the total patients examined there were 350 people whose blood maternal-side inherited genes are transmitted, which indicates the percentage of cases with genetically inherited disorders within a given population. A little higher, 400 patients inherited genetic diseases from the father, further proving that paternal link plays a major role in the occurrence of genetic diseases. Moreover, by studying the family trees of the patients, 250 patients (25%) had genetic defects related to paternal family genes, indicating that family history is important for understanding inheritance patterns and disease risk.

Relationship between birth defects and parental consent

The implication of this data is that the genetic disorder risk can no longer be dependent solely on the maternal side of the lineage. It shows that not only is it difficult to diagnose what kind of gene will be transferred, but also which couple should be examined on all possibilities in genetic screening. The study of these aspects can contribute to more precise prognostic estimates related to scantily represented hereditary diseases or facilitate the improvement of prevention and treatment strategies. Table III demonstrates the relationship between parental consent and likelihood of birth defects among the patients. These results leave one wondering how the in-

volvement of parents in the decision-making of treatment will impact congenital anomalies in children.

The group in which parental consent was given for medical treatments or interventions had 150 patients who had birth defects while 650 patients had no birth defects. This means that the health of most patients who require their parents' consent improves, could be because of early treatment and preventive measures being made when the parents agree. On the other hand, only 100 patients with birth defects were detected from the same cases that lacked parents' permission while 100 patients with no birth defects were detected. The comparison of the higher prevalence of birth defects when parents gave no consent to the procedures means that if parents deny or delay consent, premature detection and management of medical disorders may be lost opportunities, and the birth defects might worsen. These analyses stress key aspects of the informed-population model with respect to the requirement of parental consent for beneficial medical treatment and propose that informed decisions should be based on the increased risk of birth defects. Educating parents with detailed details and option about the importance of genetic test, pre-natal care maternal education, and improved medical intervention could go a long way in checking high prevalence rate of the birth defect.

Maternal health and impact on genetic disorders

The nature of relationship between the history of maternal health and the presence of genetic disorders in patients has been presented in Table IV. The data highlights two categories: families of pregnant women in poor health and families of healthy pregnant women and how these conditions may affect the chances of having children with a specific genetic disease. In the users surveyed, one hundred of the patients' mothers had complications during pregnancy, while the other fifty did not; it can therefore be concluded that the occurrence of health problems in pregnant women could increase the risk of genetic disorders among children. In contrast, among the respondents whose mothers had no serious disease, 250 had genetic diseases, and 700 had no genetic disorder. These findings imply that although maternal health may influence the development of genetic disorders, most patients without serious maternal illness had little or no impact from such diseases. However, genetic disorders were still present in a significant number of this group, which shows that there are other genes and environmental factors leading to these disorders as well. By appreciating the impact of maternal health on genetic disorder, better diagnosis during the prenatal period can be embraced as well as the practices in genetic counseling may also be improved.

Analysis of patient age vs. white blood cell count by genetic disorder

Figure 4 shows the distribution of the white blood cell count according to the patient's age and the type of genetic disorder they suffer from. Figure 5 is a star diagram, where every star represents one patient: horizontally, the age of a patient is provided in years; vertically, the WBC count is given in thousand per microliter. The plot is marked according to genetic disorders: mitochondrial genetic inheritance disorders, cystic fibrosis, and multifactorial genetic inheritance disorders. Patients with mitochondrial genetic inheritance disorders are from as young as 2 years to 12 years old with variable WBCs from low to high. Cystic fibrosis is characterized by one patient aged 4 years with mildly decreased WBC count. Multifactorial genetic inheritance disorders identified occurred in a patient age group of 6 - 11 years and the overall WBC count of the patients is approximately 7,000 - 8,000 per microliter. There are also cases with unclassified genetic disorders (NaN), usually seen in children of about 7 - 8 years with moderate white blood cell numbers. It showed that the patients' age and white blood cell counts differ across genetic disorders; this provides an observation of the patients' differential physiological susceptibility to such disorders.

Computational efficiency comparison of various approaches

The time cost (time taken in seconds) analysis for the different machine learning algorithms compared to the proposed GNN-Bi-LSTM with AM is shown in Table 5. This testimonies to the findings in the table below that

standard machine learning techniques such as the Support Vector Machine (SVM) and the Artificial Neuron Network (ANN) take a much longer time to train than their cloud counterparts. However, other algorithms with high accuracy like: K-Nearest Neighbors and Ada-Boost record a lower time taken to execute on the data, which are 24.71 seconds and 25.27 seconds respectively. However, the proposed GNN-Bi-LSTM-AM model gives better results than all other algorithms that were tested in terms of time complexity and the proposed model took only 7.85 seconds to complete. This significant amount of time saved proves the effectiveness of the proposed use of graph neural networks (GNN) for genetic relationship modeling while the Bi-LSTM layer accounts for temporal health data dependency. The addition of the attention mechanism also improves the model's performance because it directs the model to pay attention to important characteristics and generate predictions more efficiently, which makes the AI highly suitable for large scale genetic databases.

Model assessment

Measures of performance are therefore crucial in measuring outcomes of the models in application, especially in areas that are sensitive such as the medical field where accuracy can be a determinant factor of the result. Proposed GNN-Bi-LSTM with the attention mechanism is used in detecting and predicting rare genetic disorders, and the following metrics serve as the evaluation criteria: accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + F*P} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + F*n} \quad (10)$$

$$\text{F1 score} = \frac{2TP}{2TP + F*p + F*n} \quad (11)$$

The results on the proposed GNN-Bi-LSTM with attention mechanism (GNN-Bi-LSTM-AM) model to identify and prognose rare genetic disorders are summarized in Table 6. With excellent results based on diverse appraisals displayed, the model suggests how the model might address the complexity of genetic information.

The results indicate that the model has a stunning 98.5 percent accuracy, which implies that the model provides a true diagnosis and is able to detect the presence/absence of genetic disorders in many examples with a high level of certainty. At an accuracy of 96, it is clear that the model boasts of an impressive TP rate (the correctly predicted cases of genetic disorders or true positive) without multiple cases of FP or misclassification which is significant in the diagnosis of genetic disorder. The 0.96 recall rate implies that a majority of the real cases of genetic disorders are covered by the model in such a way that there are few possibilities of missing a positive case. This is especially true where, such as in the application of medicine, not diagnosing the problem can lead to illness. Confirming the model's balanced

performance, the F1-score of 97 indicates the performance given by the model for both precision as well as recall. This is evident from the high F1-score as it shows that the model is very efficient in balancing between the true positive rate and minimizing the false positives as well as false negatives in this case genetic disorders; making this model appropriate for any real-life application in healthcare especially the diagnosis of rare genetic disorders whereby both precision and recall are paramount. These metrics lend credence to the effectiveness of the GNN-Bi-LSTM-AM model proving its effectiveness, high accuracy and reliable genetic disorder prediction making the model fit for use in predictive analytics on other forms of genetic disorders.

Figure 6 represents a Receivers Operating Characteristic (ROC) curve which is a two-dimensional graphical representation of a binary classifier system that measures the true positive rate (TPR = sensitivity) against the false positive rate (FPR = 1-specificity). The curve explains the performance by the classifier in terms of positive and negative class at various thresholds. The closer the curve is to the upper left corner, the better is performance, here the classifier has a very high true positive rate compared with the false positive rate. The curve shows that the classifier has the capability of doing a good job on distinguishing between the two classes since, the closer the AUC value to 1, then the better is the classifier on its job of classification. The line at the bottom represents a random guess (AUC at 0.5) while the classifier curve extends far beyond this, denoting a good amount of predictive power.

Figure 7 compares the performance of several machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, and the proposed GNN-Bi-LSTM with Attention Mechanism (AM), across key performance metrics: include level of accuracy, precision, recall and F1-score or else known as measure of statistical goodness. SVM performs well with accuracy of 90%, precision of 89 % recall of 90% and F1-score of 89.5%. Although they are good enough these overtake SVM and KNN and Naïve Bayes yielded a better result, out of 1,000 samples of data, KNN, had an accuracy of 94% for the test data with precision of 92% and recall of 91% with an F1 of 92.5/1,000 which shows that the KNN algorithm has a better balance of returns between precision of correct prediction and recalling KNN and Naïve Bayes have high efficiency in the probabilistic modelling of genetic disorder prediction with an accuracy of 95%, precision of 93%, recall of 92%, and an F1-score of 93%.

The proposed GNN-Bi-LSTM-AM model performs remarkably well as compared to the other methods developed so far, with an accuracy of 98.5%, precision of 96%, recall of 96%, and F1-score of 97%. Superior performance can be explained by the fact that GNN enables the model to capture complex genetic relations, Bi-LSTM allows us to handle sequential data representing health and, finally, the attention mechanism helps to emphasize the most important features. The results

show that the proposed model offers a more accurate prediction model and better balance between precision and recall, thus making for a perfect solution for the live identification and prediction of the rare genetic disorders (Table 7).

CONCLUSION

The integration of Graph Neural Networks (GNNs) with Bidirectional Long Short-Term Memory (Bi-LSTM) and attention mechanisms demonstrated outstanding performance in predicting rare genetic disorders. Using 228 cross-validated samples, the model achieved 98.5% accuracy with 96% precision and recall, effectively identifying individuals at genetic risk. GNNs captured intrinsic genetic relationships, while Bi-LSTM enhanced sequential analysis of physiological data such as heart and respiratory rates. These results highlight the potential of advanced machine learning for early and accurate diagnosis of genetic disorders, providing a robust foundation for future research, larger datasets, and improved interpretability in personalized genomic medicine. The further work will focus on the extensive validation of the suggested framework of GNN-Bi-LSTM-AM exploiting the independent and clinically verified datasets derived out of various genetic and clinical repositories. The external validation will help to verify that the model is general, robust, and can be repeated in other real-world scenarios and also allow performing the evaluation of its performance in a heterogeneous population to determine whether it is reliable in the actual diagnostic conditions. Despite the excellent predictive accuracy on the part of the framework, some limitations have to be noted such as the possibility of overfitting as well as limited diversity of the dataset due to the use of the open-source genetic data. By including data of larger sizes, and multi-institutional, and clinically validated data, these issues will be resolved and will further increase the scientific rigor and applicability of the model. Besides, further studies are needed to examine how the multi-omics layers (transcriptomic, proteomic, metabolomic, epigenomic and so on) can be integrated to obtain a more detailed biological picture and demonstrate the intricate interactions between genes and phenotypes. All these improvements will contribute to the increased diagnostic accuracy, the increase in the translational applicability of the model, and its wider adoption in precision medicine and genomic healthcare.

Source of Funds:

This Study received no specific grants from public, commercial, or not-for-profit funding agencies.

Use of AI-Assisted Technology:

The authors have carefully reviewed and verified all AI-generated content to ensure accuracy, originality, uni-

queness, and compliance with ethical and academic standards.

Declaration of Interest:

The authors do not have any conflict of interest.

References:

1. Schaaf J, Sedlmayr M, Schaefer J, Storf H. Diagnosis of rare diseases: a scoping review of clinical decision support systems. *Orphanet J Rare Dis* 2020;15(1):263. (PMID: 32972444)
2. Frederiksen SD, Avramović V, Maroille T, Lehman A, Arbour L, Tarailo-Graovac M. Rare disorders have many faces: in silico characterization of rare disorder spectrum. *Orphanet J Rare Dis* 2022;17(1):76. (PMID: 35193637)
3. Jacquemont S, Huguet G, Klein M, et al. Genes to Mental Health (G2MH): a framework to map the combined effects of rare and common variants on dimensions of cognition and psychopathology. *Am J Psychiatry* 2022;179(3):189-203. (PMID: 35236119)
4. Mbemi A, Khanna S, Njiki S, Yedjou CG, Tchounwou PB. Impact of gene-environment interactions on cancer development. *Int J Environ Res Public Health* 2020;17(21):8089. (PMID: 33153024)
5. Groft S, De la Paz M, Taruscio D. Progress, challenges, and global approaches to rare diseases. *Acta Paediatr* 2021;110(10):2711-6. (PMID: 34105798)
6. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 2023;23(1):689. (PMID: 37740191)
7. Singh AV, Chandrasekar V, Paudel N, et al. Integrative toxicogenomics: advancing precision medicine and toxicology through artificial intelligence and OMICs technology. *Biomed Pharmacother* 2023;163:114784. (PMID: 37121152)
8. Onnekink AM, Klatte DCF, van Hooft JE, et al. Attitudes toward genetic testing, family planning and preimplantation genetic testing in families with a germline CDKN2A pathogenic variant. *Fam Cancer* 2024;23(3):255-65. (PMID: 38822936)
9. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput Biol Med* 2022;149:106043. (PMID: 36115302)
10. Zhao AP, Li S, Cao Z, et al. AI for science: predicting infectious diseases. *J Saf Sci Resilience* 2024;5(2):130-46. <https://www.sciencedirect.com/science/article/pii/S266644962400015X>
11. Diao N, Abed M, Taha S, Ali M. Machine learning and traditional statistics integrative approaches for bioinformatics. *J Ecohumanism* 2024;3:335-52. https://www.researchgate.net/publication/383772172_Machine_Learning_and_Traditional_Statistics_Integrative_Approaches_for_Bioinformatics
12. Nandhini K, Tamilpavai G. An optimal stacked ResNet-BiLSTM-based accurate detection and classification of genetic disorders. *Neural Process Lett* 2023;55(7):9117-38. (PMID: 37359129)
13. Xu Y, Vuckovic D, Ritchie SC, et al. Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genomics* 2022;2(1):None. (PMID: 35072137)
14. Rahman A, Nasir M, Gollapalli M, et al. Advance genome disorder prediction model empowered with deep learning. *IEEE Access* 2022;10:70317-28. <https://ieeexplore.ieee.org/document/9810256>
15. Rahman A, Nasir M, Gollapalli M, et al. IoMT-based mitochondrial and multifactorial genetic inheritance disorder prediction using machine learning. *Comput Intell Neurosci* 2022;2022(1):2650742. (PMID: 35909844)
16. Arloth J, Eraslan G, Andlauer T, et al. DeepWAS: multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLoS Comput Biol* 2020;16(2):e1007616. (PMID: 32012148)
17. Aromolaran O, Beder T, Oswald M, Oyelade J, Adebisi E, Koenig R. Essential gene prediction in *Drosophila melanogaster* using machine learning approaches based on sequence and functional features. *Comput Struct Biotechnol J* 2020;18:612-21. (PMID: 32257045)
18. Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. *Nat Rev Genet* 2021;22(5):269-83. (PMID: 33408383)
19. Genomes and genetic disorder / prediction II [Internet]. Kaggle; 2024 [cited 2024 Oct 24]. Available from: <https://kaggle.com/code/brsdincer/genomes-and-genetics-disorder-prediction-ii>
20. Acharya DB, Zhang H. Feature selection and extraction for graph neural networks. *arXiv* 2019 Oct 25. <https://arxiv.org/abs/1910.10682>