

Explainability Aware Heterogeneous Temporal Graph Transformer for Automated Genetic Disorder Detection from Genomic Data

Puvaneswari Subramaniapillai¹, Indirani Govindasamy² and Revathy Gurumurthy³

¹Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, Tamilnadu, India

²Department of Computer Science and Engineering, Government College of Engineering, Sengipatti, Thanjavur, Tamilnadu, India

³Department of Computer Science and Engineering, SASTRA Deemed University, Kumbakonam, Tamilnadu, India

Article history

Received: 15-05-2026

Revised: 07-06-2026

Accepted: 23-09-2026

Corresponding Author:

Puvaneswari Subramaniapillai
Department of Computer
Science and Engineering,
FEAT, Annamalai University,
Chidambaram, Tamilnadu,
India

Email: bhuvanacke2021@gmail.com

Abstract: Genetic disorders affect millions worldwide, posing major challenges for accurate diagnosis and effective treatments. New developments in genomic technologies, especially Next-Generation Sequencing (NGS), have enhanced the classification of genetic modification and improved diagnostic abilities. Moreover, the Machine Learning (ML) and Deep Learning (DL) methods emerged as efficient tools for evaluating the intricate genomic data and determining the patterns, which can be generally complex to identify by employing conventional analytical techniques. Additionally, the Explainable Artificial Intelligence (XAI) has developed as a significant constituent in the analysis of genomic data by enhancing the interpretability of DL methods and showing the contribution of main genetic features involved in the detection of genetic disorders. This study presents an Explainability-Aware Heterogeneous Temporal Graph Transformer for Automated Genetic Disorder Detection (XHTGT-AGDD) approach. The primary objective of the proposed model is to improve the reliability and accuracy of genetic disorder detection from genomic data by integrating advanced techniques. Initially, preprocessing is performed using quantile transformation to normalize the input data for robust analysis. Following preprocessing, feature selection is performed using the fire hawk optimizer to identify the most informative feature subsets from the genomic data. Besides, the classification task is accomplished using a heterogeneous temporal graph transformer, which captures complex relationships and temporal dependencies among genomic features. To further refine the learning process, hippopotamus optimization algorithm is employed for model optimization. In addition, an explainability mechanism based on SHAP is incorporated to interpret the model predictions and highlight the individual genetic features toward the final decision. The experimental result analysis of the XHTGT-AGDD model is performed using a benchmark dataset. The simulation results reported the enhanced results of the XHTGT-AGDD technique in terms of several measures, achieving an accuracy of 99.12%. Therefore, the proposed model provides reliable and effective performance for automated genetic disorder detection from genomic data.

Keywords: Genetic Disorder, Genomic Data Analysis, Feature Selection, Heterogeneous Temporal Graph Transformer, Explainable Artificial Intelligence

Introduction

Genetic disorders stem from DNA sequence alterations or structural chromosomal damage. Since the

genome comprises the genetic information, any modifications in the genome may cause variations in the structural or functional characteristics of an organism (Srinivas et al., 2025). As functional units comprised of

DNA, genes can lead to hereditary diseases when their sequence is mutated. The genome data holds significant information and medical indicators, which are employed for analyzing the genetic disorders (Gélin et al., 2026). Numerous genetic disorders, namely single gene inheritance disorders, chromosomal disorders, and complex disorders, have been analyzed corresponding with the structures of DNA (Raj et al., 2025). The single gene disorder results from a structural change in a single gene. While, chromosomal aberrations, namely extra, missing, or rearranged segments, lead to genetic disorders. The next type of complex disease has been triggered by a multigenic alterations (Xue et al., 2026). With the fast development of advanced biomedical technologies, massive amounts of biomedical information are produced that contain medical images, genomic sequences, disease-based SNPs, and gene expression profiles (Shahid et al., 2026). This data offers a great, useful possibility to comprehend the disease mechanisms and increase the disease diagnosis and treatments at the molecular level. Various categories of biomedical data give disease-based information through different biological and medical conditions (Kirchhoff et al., 2025). Thus, numerous computational techniques are established for analyzing and extracting useful information from biomedical data.

Recently, the usage of Machine Learning (ML) and Deep Learning (DL)-driven methodologies has offered a remarkable achievement in various domains, such as automation, prediction, medicine, prognosis, and more. These methods have been trained using better-quality previous data (Guo et al., 2025). The ML or DL methods determine the patterns or correlations in the data to create the prediction results. The DL methods, especially Neural Networks (NNs) with numerous Hidden Layers (HLs), are proficient in learning intricate and higher-dimensional data models (Minten et al., 2025). The aforementioned systems provide automatic predictions and execute assistive parts for medical specialists, thereby highlighting the sensitivity and significance of these tasks. The selection of ML and DL methods is reliant on the form of real-time complex. These approaches are a huge diversity of possible applications in the bioinformatics domain (Jurgens et al., 2025). The rapid development of biological data increases the complexity of data management. The ML and DL methods have been implemented for making predictive decisions, which depend on great gene sequences, and to handle the large-scale biological information (Raji et al., 2025). Several biological fields are presently utilizing such techniques for feature engineering. The applications comprise evaluations of X-rays, genome-wide relationships, prediction of enzyme functions and protein functions, and so on.

Novelty and Contributions

Many Machine Learning (ML) and Deep Learning

(DL) methods have been proposed for genetic disorder prediction, most current works focus on conventional neural networks, ensemble learning or homogeneous graph-based architectures which cannot fully account for the complex interactions between genomic attributes. Moreover, there has been little focus on incorporating explainability and optimization components into a single system for trustworthy clinical decision making.

In order to address these issues, this study introduces the Explainability-Aware Heterogeneous Temporal Graph Transformer for Automated Genetic Disorder Detection (XHTGT-AGDD). The suggested framework integrates feature normalization, metaheuristic feature selection, heterogeneous graph-based representation learning, parameter optimization, and XAI within one framework. The following are the key contributions of this work:

- Employ Fire Hawk Optimizer (FHO) for feature selection, identifying the most informative feature subsets from genomic data
- Introduce a Heterogeneous Temporal Graph Transformer (HTGT) model, which enables the model to effectively learn structured patterns present within genomic data and improves the accuracy of disorder classification
- Integrate the Hippopotamus Optimization (HO) for model optimization and parameter tuning, resulting in improved classification performance

Related Studies on Genetic Disorder Detection

This section presents a brief synthesis of the literature concerning the genetic disorder detection. Puvaneswari and Indirani (2026) developed a deep Feature Selection (FS) mechanism for genetic disorder detection by using the convolutional autoencoder and metaheuristic optimization technique, aiming to design a precise DNA-enabled genetic disorder detection scheme employing innovative methods for earlier and dependable disease diagnosis. Shahin et al. (2025) aimed to develop a computational gene system, with a specific focus on Autosomal Recessive (AR) diseases. The model intended to design a hybrid Convolutional Neural Network-Gated Recurrent Unit (CNN-GRU) architecture, particularly established for identifying the genetic diseases widespread in Saudi Arabia. Keerthana and Mahesh (2025) developed current developments in ML-based prediction models of genetic disorders. This can be highlighted by the standard ML methods, the FS approaches, the DL framework, and the multi-omics data integration methods. In addition, the complexities relevant to computational resource necessities, data imbalance, ethical concerns, and model interpretability have been considered. This study demonstrates that ML-based genomic prediction has enormous potential for

advanced healthcare. Chen et al. (2025a) implemented the medical applications of AI based analytics in leveraging tailored information for autoimmune rheumatic diseases. By using the synergistic method of incorporating AI over different data sets, the medical experts have gained comprehensive views of patient health and possible risks. The ML methods are used to identify the higher-risk patients, predict disease activities, and enhance therapeutic approaches, which can depend on medical, immunological, and genomic profiles.

Gwak et al. (2024) inspected the pervasiveness of genetic variations and the number of alternative myocardiopathies in patients with suspected HCM in Korea, and compared the results of HCM with respect to the existing sarcomere gene mutation. The genetic analysis has been executed by employing a next-generation sequencing panel. Olivucci et al. (2024) examined the LRS application for diagnosing the morphological forms variations and replication of this growths, possibly due to the identification methods have not progressed as quickly as those devoted to Single Nucleotide Variants (SNV) analysis: Benchmark methods are microarrays and karyotyping for unbalanced and balanced chromosome mutations for the strengthening and sizing of extended alleles, influenced by restricted resolutions and sensitivity that the development of NGS cannot majorly increase. Alzahrani and Alharithi (2024) developed an innovative approach that utilizes the Random Forest (RF) data Whole-Genome Sequencing (WGS) and evaluation to address these limits. The WGS offers an all-inclusive genetic profile of patients, thus gaining an extensive range of genetic mutations. With the help of the RF technique, it surpasses in managing the difficult datasets and identifying the non-linear interactions. This aforementioned method is aimed at uncovering complex relationships among gene variations and rare genetic diseases.

Proposed Model Architecture

This study introduces the XHTGT-AGDD model, a structured approach for identifying and classifying genetic diseases from genomic datasets. The process begins with data preprocessing using quantile normalization to enhance input quality. Next, the Fire Hawk Optimizer (FHO) is employed to select the most relevant, high-impact features. Following this feature selection, the Heterogeneous Temporal Graph Transformer (HTGT) model is applied for precise classification. The model then optimizes hyperparameter settings using the Hippopotamus Optimization (HO) algorithm, significantly improving model stability and performance. Lastly, SHAP technique are incorporated to ensure transparency and provide detailed breakdown of the diagnostic rationale. Fig. 1 demonstrates the overall workflow of the proposed approach.

Genomic Data Pre-Processing

The genomic dataset is pre-processed using Quantile Transformation (QT). The QT process includes ranking the features of every sample through the magnitude, computing the average values for features, gaining the same rank, and then replacing the values of the entire features that are occupied with these average values (Zhao et al., 2020). The following stage is to rearrange the features of every instance in their unique sequences. This sequences of stages are represented by the QT and permits the alteration of the dataset into a uniform distribution over the instances. Among the QT approaches, the “All” approaches normalize the data as the comprehensive types of irrespective of class factors, in which whole instances have been altered to keep a common feature distribution. This method supports decreasing the chances that are affected by variances in feature measurements throughout the analysis. The “Class-specific” technique is separated into the dataset with respect to the class labels, where every individual subset of the class is QT self-reliantly. By using such measures, the QT efficiently minimizes distributional differences among samples and stabilizes the data, thus increasing the learning ability of the classification method for the identification of genetic disorders:

$$X = \{x_{ij}\}, i = 1, 2, \dots, n, j = 1, 2, \dots, m \quad (1)$$

Here, the parameter n indicates the no. of instances, and m represents the no. of attributes in the genetic dataset.

Metaheuristic Based Feature Selection

Relevant genomic features are selected using the FHO algorithm for identifying the most informative feature subsets. FHO is an innovative optimization model intended to tackle intricate, non-convex, and non-linear optimization issues (Moghaddam et al., 2025). Motivated by the usual behavior of fire hawks, while they carry burning sticks to unburnt regions, causes prey to arise that they search for a unique strategy that permits the FHO to excel in surroundings necessitating adaptable hunting mechanisms and stochastic movement to discover large solution areas. The primary phase in the FHO comprises an arbitrary position inside the searching area, where the location of every fire hawk X_{ij} can be specified:

$$X_{ij} = rand \times (U_j - L_j) + L_j, j = 1, 2, \dots, D \quad (2)$$

Now, $rand$ indicates an arbitrary number, and L_j and U_j depict the low and up bounds of the searching domain. Once the initial location can be determined, every solution X can be assessed, and the finest solution, X_b , can be recognized depending on fitness criteria originating from every flame:

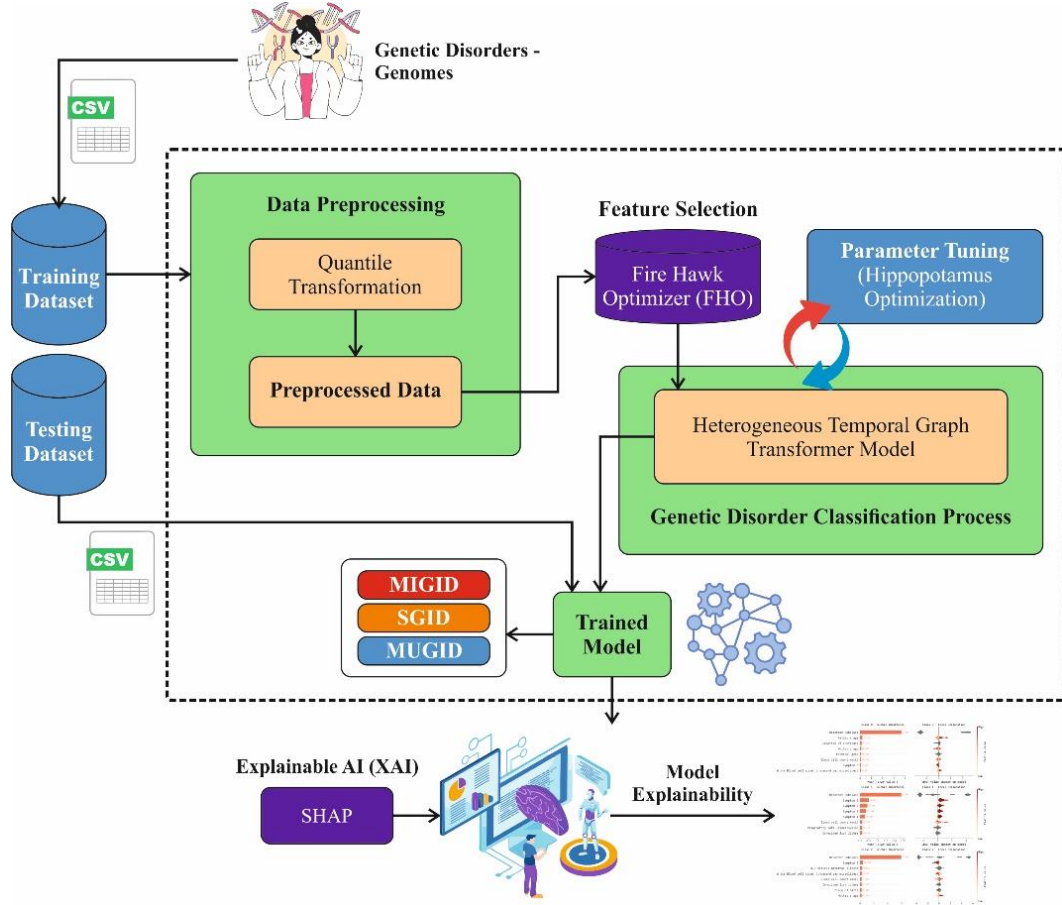


Fig. 1: Overall workflow of the proposed approach

$$D_{lk} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, l = 1, 2, \dots, m \quad (3)$$

The location of FH has been upgraded depending on the finest solution and arbitrary movement parameters:

$$FH_l(t+1) = FH_l(t) + (r_1 \times X_b - r_2 \times FH_n(t)), l = 1, 2, \dots, n \quad (4)$$

Now, $FH_n(t)$ depicts the preceding location of Fire Hawk's, r_1 and r_2 refer to random values induced by the magnitude of movement:

$$SP_l = \frac{\sum_{q=1}^r PR_q}{r}, q = 1, 2, \dots, r; l = 1, 2, \dots, n \quad (5)$$

The movement of prey is upgraded:

$$PR_q(t+1) = PR_q(t) + (r_3 \times FH_l - r_4 \times SP_l(t)), l = 1, 2, \dots, n; q = 1, 2, \dots, r \quad (6)$$

The fire hawks endure to search by altering their location, whereas prey continuously upgrade their position:

$$P_l = \frac{\sum_{k=1}^m PR_k}{r}, k = 1, 2, \dots, m \quad (7)$$

Afterwards, the prey also upgraded its location:

$$PR_q(t+1) = PR_q(t) + (r_5 \times FH_a - r_6 \times SP(t)), l = 1, 2, \dots, n; q = 1, 2, \dots, r \quad (8)$$

To optimize feature selection, the proposed Fitness Function (FF) balances higher classification effectiveness with a minimal feature subset, described by Eq. (9):

$$Fitness = \alpha \gamma_R(D) + \beta \frac{|R|}{|C|} \quad (9)$$

Now, $\gamma_R(D)$ defines the misclassification rate of the given model. $|R|$ means chosen subset cardinality, and $|C|$ indicates complete feature counts; α and β are 2 metrics linking subset size to predictive performance.

Deep Graph Transformer Classification

The Heterogeneous Temporal Graph Transformer (HTGT) is chosen because of its ability to model both long-range context and heterogeneous feature

interactions. The recurrent models like LSTM and GRU mainly focus on capturing the sequential dependencies while traditional CNN-based approaches are more concerned with local feature extraction. In the same way, traditional graph neural networks like GraphSAGE and GCN integrate neighborhood information but struggle to effectively capture various types of features and long-range interactions. HTGT, on the other hand, integrates the graph representation learning with the transformer self-attention mechanisms. The use of a heterogeneous graph structure allows the effective modelling of different kinds of feature relationships, and the attention mechanism models local and global dependencies among the genomic attributes. As a result, HTGT is more expressive in terms of feature representations and more powerful in terms of classification than traditional graph-based and sequence-based approaches. The selected features are classified using an HTGT architecture to capture complex relationships among genomic features.

Graph Embedding Layer: The conventional transformer architecture is extremely efficient to handle the sequential and structural biological data; consequently, the preliminary stage is used to change from the genetic data under the graph representation to a proper input representation (Wang, 2025). Present Graph Transformers directly employ the entire graph for producing input sequences. Although the large-scale type of genetic feature graphs creates this method inappropriate, as the wide-ranging feature correlations and node interconnection marginalize the local-view architecture of nodes and determine the higher computational expenses. In the graph embedding layer, especially the genetic feature at the node form v , the graph convolution-based aggregation to separately integrate the various categories under every data sample as multi-perspective lower-views of structural data. This method must be represented as given below:

$$H_{v,r}^t = A_r^t H_{N_r^t(v)}^t, \quad (10)$$

Here, the mathematical term $H_{v,r}^t$ represents the adjacent relation r , feature node type v with instance index t , $N_r^t(\cdot)$ means node type under relation r with respect to instance index t , A_r^t remains the normalization of the adjacency matrix according to the relation r . In conclusion, after that integrating entire forms of neighbors distinctly, and gathered a lower-views structural data and and genetic feature representation (input sequence) for the Transformer $\{H_{v,r}^t | r \in \mathcal{R}(v), t \in T\} \cup \{H_v^t | t \in T\}$, Now, the parameter $R(v)$ indicates the sets of relationships with v and target node types, T represents the correct of data samples. It can be decreased the input sequences, length of the transformer from the overall numbers of nodes ($|V|$) to $(T \cdot |T_n|)$. Here, the mathematical term T_n is the set of node categories.

Hetero-Temporal Encoder: The transformer model only gains the feature-based resemblance among the nodes; it is prone to variable feature disseminations in genetic disorder datasets. Motivated by original transformers that utilize the positional encoders for correcting the attention biases influenced through comparative positions for improving the Self-Attention Mechanism (SAM). The hetero-temporal encoder is developed for genetic feature graphs. In a specified node type v , primarily design the representation of the node type that comprises descriptive encoding of feature categories. The embeddings are gained from this procedure and indicate the semantic feature node types, and then enable the extraction of heterogeneous data using the genetic feature interrelations. The temporal encoding is used as a positional representation mechanism to maintain a sense of ordering in the feature representations produced in the construction of the graph. This positional encoding, combined with the sinusoidal features, helps to add context to the information, allowing the transformer to capture the relationships between features in the learned graph representation. This method must be expressed by:

$$\text{Prompt}(v) = \{\text{Introduction to type } v; \text{Instruction.}\}, \quad (11)$$

$$H_v^{LLM} = LLM(\text{Prompt}(v)), \forall v \in T_n \quad (12)$$

Here, the parameter H_v^{LLM} indicates the type encoding vector of type v amplified by semantic features. For sequential structures of feature representations, the integration of time-dependent sinusoidal functions for encoding the sequential positional information into the H_v^{LLM} vector as given below:

$$p_i^t = \begin{cases} \sin(t/10000^{2i/d}) & \text{if } i = 2k \\ \cos(t/10000^{2i/d}) & \text{if } i = 2k + 1 \end{cases} \quad (13)$$

$$H_v^{sp,t} = ||_{i=1}^d (H_v^{LLM} e + p_i^t), \quad (14)$$

Whereas the mathematical form d indicates the dimension of the vector, i refers to the index of elements, and $H_v^{(sp,t)}$ represents the spatio-temporal encoder process of feature node type v with respect to instance index t . Fig. 2 shows the architecture of the Heterogeneous Temporal Graph Transformer (HTGT) model.

Then, the concatenation of the spatio-temporal encoding $H_v^{(sp,t)}$ with the genetic feature representations H_v^t in accordance with the type and index as given below:

$$Z_v^t = \text{concat}(H_v^t, H_v^{(sp,t)}) \quad (15)$$

In conclusion, the original input sequence can be enhanced for successive processing and derived by:

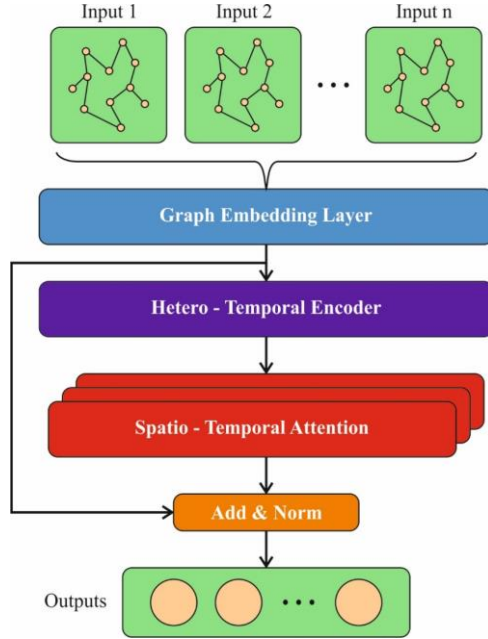


Fig. 2: Architecture of the HTGT model

$$Z = \{Z_{(v,r)}^t \mid r \in R(v), t \in T\} \cup \{Z_v^t \mid t \in T\} \quad (16)$$

Spatio-Temporal Attention: This model is developed to combine the feature interaction data to predict the class representations of genetic disorders. With the usage of the hetero-temporal encoder to accurately estimate the bias of feature-based attention in diverse genetic feature graphs, various forms of nodes should be exchanged in the training samples. Precisely, the specified target node v , the dot-product attention mechanism (AM) from the conventional transformer, as represented by using these mathematical forms:

$$Q = W_Q Z_v, K = W_K Z, V = W_V Z, \quad (17)$$

$$Z'_v = \text{Softmax} \frac{QK^T}{\sqrt{d_K}} V, \quad (18)$$

Here, the mathematical terms W_K, W_Q, W_V describe trainable samples, $Z_v = [Z_v^1, \dots, Z_v^T]$ indicate improved target feature representations, and $Z'_v = [Z_v^{(T+1)}, \dots, Z_v^{2T}]$ refer to predictable feature representations employed for classifying the genetic disorder. Lastly, it is flattened and mapped for the genetic disorder prediction layer by employing the linear layer.

Model Parameter Tuning

The model parameters are optimized using the HO algorithm to enhance the classification performance. The HO model is motivated by the behavior of hippopotamuses (Chen et al., 2025b).

Initialization: Alike a conventional swarm intelligence optimizer model, the initialization procedure of this methodology is:

$$X_{i,j} = Y_{i,j} + r(Y_{u,j} - Y_{l,j}) \quad i = 1, 2, \dots, N; j = 1, 2, \dots, M \quad (19)$$

Now M indicates decision variable counts for the problem, N reflects the demographic size, and j denotes the existing decision variable, and $X_{i,j}$ indicates the candidate position. $Y_{u,j}$ and $Y_{l,j}$ refer to the lower and upper limits, and r depicts an arbitrary number.

Position Update of Hippos: The population of hippo comprises adult females, juvenile hippos, various adult male hippos, and a prevailing male hippo:

$$X_{i,j}^{Mhippo} = X_{i,j} + r(X^{Dhippo} - I_1 X_{i,j}) \quad i = 1, 2, \dots, \frac{N}{2}, j = 1, 2, \dots, M \quad (20)$$

$$S = \begin{cases} I_2 \times r_1 + (\sim p_1) \\ 2 \times r_2 - 1 \\ r_3 \\ I_1 \times r_4 + (\sim p_2) \\ r_5 \end{cases} \quad (21)$$

Now, $X_{i,j}^{Dhippo}$ marks where the top-performing hippo is located in this round, $X_{i,j}^{Mhippo}$ depicts the position of male hippo, I_1, I_2 refers to integer with the range of $[1, 2]$, S represents 5 arbitrarily chosen scenarios, $\sim p_1, \sim p_2$ indicates either 0 or 1, chosen arbitrarily, r_5 depicts a randomly generated number and $r_1 \sim r_4$ refers to an arbitrary vector in range of $[0, 1]$:

$$T \exp\left(-\frac{t}{T_{\max}}\right) \quad (22)$$

$$X_{i,j}^{FBhippo} = \begin{cases} X_{i,j} + S_1(X^{Dhippo} - I_2 h_i)T > 0.6 \\ Z \text{ else} \end{cases} \quad (23)$$

$$Z = \begin{cases} X_{i,j} + S_2(h_i - X_{i,j}^{Dhippo})r_6 > 0.5 \\ Y_{i,j} + r_7(Y_{u,j} - Y_{l,j}) \text{ else} \end{cases} \quad (24)$$

Now, t denotes the existing iteration number, $T_{\max}$ indicates the maximal iteration number, and T is the count of remaining iterations. h_i represents the average location of some randomly chosen hippos, $X_{i,j}^{FBhippo}$ refers to the position of an immature hippo, S_1, S_2 refers to the state of 5 randomly chosen scenarios, Z represents the status, and r_7 denotes an arbitrary number. The position upgrade equation for female, immature, or male hippos:

$$X_i = \begin{cases} X_i^{Mhippo} & F_i^{Mhippo} < F_i \\ X_i & \text{else} \end{cases} \quad (25)$$

$$X_i = \begin{cases} X_i^{FBhippo} & F_i^{FBhippo} < F_i \\ X_i & \text{else} \end{cases} \quad (26)$$

Now, F^{hippo} indicates male hippo's objective value, F^{Bhippo} denotes the target function outcome for non-dominant or young hippos, and F_i depicts the value of the objective function.

Exploration Phase: Owing to their huge size, can effectually preclude predators from approaching. Still, young, curious hippo leave the bloot:

$$X_j^{Predator} = Y_{i,j} + r_8(Y_{u,j} - Y_{l,j}) \quad j = 1, 2 \dots M \quad (27)$$

$$D = |X_j^{Predator} - X_{i,j}| \quad (28)$$

Now, D indicates the gap separating the hippo from its threat, and r_8 indicates an arbitrary vector. $F_j^{Predator}$ depicts the target value of the hippo in the defense exploration stage:

$$X_{i,j}^{HippoR} = \begin{cases} RL \oplus X_j^{Predator} + \frac{\alpha}{c-d \times \cos(2\pi\beta)} \times \frac{1}{D} F_j^{Predator} < F_i \\ RL \oplus X_j^{Predator} + \frac{\alpha}{c-d \times \cos(2\pi\beta)} \times \frac{1}{2 \times D + r_9} F_j^{Predator} \geq F_i \end{cases} \quad (29)$$

Where $\oplus$ indicates the result of the XOR operation, and $X_{i,j}^{HippoR}$ depicts the position of the hippo facing the predator. The Levy distribution is described:

$$Levy(\vartheta) = 0.05 \times \frac{\omega \times \sigma_\omega}{|\mu|^{\frac{1}{\vartheta}}} \quad (30)$$

$$\sigma_\omega = \left[\frac{\Gamma(1+\vartheta) \sin(\frac{\pi\vartheta}{2})}{\Gamma(\frac{1+\vartheta}{2}) \Gamma(\frac{\vartheta-1}{2})} \right]^{\frac{1}{\vartheta}} \quad (31)$$

Now, ϑ indicates a constant equal to 1.5, Γ indicates the Gamma function, ω and μ refer to arbitrary numbers in the interval of [0,1], and σ_ω depicts the Levy step size:

$$X_i = \begin{cases} X_i^{HippoR} & F_i^{HippoR} < F_i \\ X_i & F_i^{HippoR} \geq F_i \end{cases} \quad (32)$$

Eq. (28) indicates the last location of hippo at that phase, and F_i^{HippoR} depicts the target value of hippo at the last place. While $F_i^{HippoR} \geq F_j$, it indicates that the hippo was killed and substituted by another hippopotamus.

Exploitation Phase: While challenged by many predators or unsuccessful defense endeavors against them:

$$\begin{cases} Y_{i,j}^{local} = \frac{Y_{i,j}}{t} \\ Y_{u,j}^{local} = \frac{Y_{u,j}}{t} \end{cases} \quad t = 1, 2, \dots, T_{max} \quad (33)$$

$$X_{i,j}^{HippoE} = X_{i,j} + r_{10}[Y_{i,j}^{local} + K(Y_{u,j}^{local} - Y_{i,j}^{local})] \quad (34)$$

Where $Y_{u,j}^{local}$, $Y_{i,j}^{local}$ depict the upper and lower limits of the existing safer place, $X_{i,j}^{HippoE}$ refers to hippo's position for neighboring secure location, K denotes an arbitrarily chosen value from 3 scenarios, and r_{10} indicates a randomly generated number in the interval of [0,1]:

$$= \begin{cases} 2 \times r_{11} - 1 \\ r_{12} \\ r_3 \end{cases} \quad (35)$$

Now, r_{12} indicates a value generated from a distribution, r_{13} denotes a uniformly distributed random real number, and r_{11} indicates a random vector in the range of [0,1]:

$$X_i = \begin{cases} X_{i,j}^{HippoE} & F_i^{HippoE} < F_i \\ X_i & F_i^{HippoE} \geq F_i \end{cases} \quad (36)$$

Now F_i^{HippoE} indicates the objective value.

The fitness choice is a significant factor manipulating the performance of the HO model. The algorithm includes a mechanism for encoding parameter sets to measure the fitness of candidate solution during optimization. The HO algorithm utilizes precision as the key standard for the FF design, expressed as follows:

$$Fitness = \max(P) \quad (37)$$

$$P = \frac{TP}{TP + FP} \quad (38)$$

Here, TP and FP denote the true and the false positive values.

SHAP-driven Model Explanation

The SHAP technique is used for interpreting the model predictions and analyzing the contribution of individual genomic features. SHAP is a game-theoretic method utilized for explaining the outcome of the DL techniques by assigning all features a significant value for providing a prediction to take the average marginal contribution (Park et al., 2026). The method assesses each possible integration of features by considering the inclusion and exclusion of a specific feature in the process of prediction and measures how the technique's outcome changes. Shapley values, originally developed in cooperative game theory, are employed for determining the contribution of all participants in the coalition game. This approach assesses a player's varied impact by quantifying their marginal contribution. Afterward, the average of the distinct contributions has been calculated, as shown in Eq. (39) as follows:

$$(\phi_i(f, x^i) = \sum_{S \subseteq N} \frac{|S|!(M-|S|-1)!}{M!} \star [f_{(S \cup \{i\})} - f_{(S)}]) \quad (39)$$

Here, M indicates the total number of features, and the summation is performed over all possible subsets S that do not include feature i . The function f implies the prediction model evaluated through the feature subset S . However, the exact computation of Shapley values is computationally expensive as it necessitates evaluating the contribution of a feature across numerous feature combinations.

Results and Performance Analysis

The performance of the proposed model is assessed through experimental analysis using the Predict the genetic disorders dataset (Kaggle, 2026). Genetic disorders are caused by DNA mutations or chromosomal

changes. Many common diseases are hereditary, and genetic testing helps in prevention, early detection, and treatment. With population growth, the incidence of these disorders is increasing, partly owing to low awareness of genetic testing. Testing during pregnancy is crucial. Out of 43 features, 24 have been selected. The dataset contains 18962 records under four classes as exposed in Table 1. There are three main categories of genetic disorders: those of Mitochondrial Genetic Inheritance (MIGID), those of Single Gene Inheritance (SGID), and those of Multifactorial Genetic Inheritance (MUGID). The data set was split into a training set and a test set in the ratio of 7:3. About 70% of the samples were used to train the model and the other 30% to test the model independently.

Python is used to implement the proposed XHTGT-AGDD framework and run it on a workstation with GPU acceleration. The Adam optimization algorithm was used in the training process to update the weights and the Hippopotamus Optimization algorithm was used for the tuning process. Model was trained for 100 epochs to ensure convergence and stable learning with an appropriate batch size. Tables 2 and 3, shows hyperparameter configuration and computational environment settings.

The selected hyperparameter configuration was determined through iterative optimization using the HO algorithm to maximize classification performance.

Fig. 3 presents the classifier results of the XHTGT-AGDD technique at 70% TRAP and 30% TESP. Figs. 3a-3b show the confusion matrix with proficient classification of each class. Fig. 3c displays the PR analysis, indicating

maximum performance across all classes. Lastly, Fig. 3d illustrates the ROC investigation, demonstrating proficient results with optimal ROC values for different classes.

Table 1: Dataset details

Class Labels	Description	No. of Samples
“MIGID”	“Mitochondrial genetic inheritance disorders”	9686
“SGID”	“Single-gene inheritance diseases”	7291
“MUGID”	“Multifactorial genetic inheritance disorders”	1985
Total No. of Samples		18962

Table 2: Hyperparameter configuration

Hyperparameter	Value
Learning rate	0.001
Batch size	32
Epochs	100
Hidden dimension	128
Transformer layers	4
Attention heads	8
Dropout rate	0.20
Optimizer	Adam
Activation function	ReLU

Table 3: Computational environment

Component	Specification
Processor	Intel Core i7
RAM	16 GB
GPU	NVIDIA RTX Series
Programming language	Python
Deep learning framework	TensorFlow/PyTorch

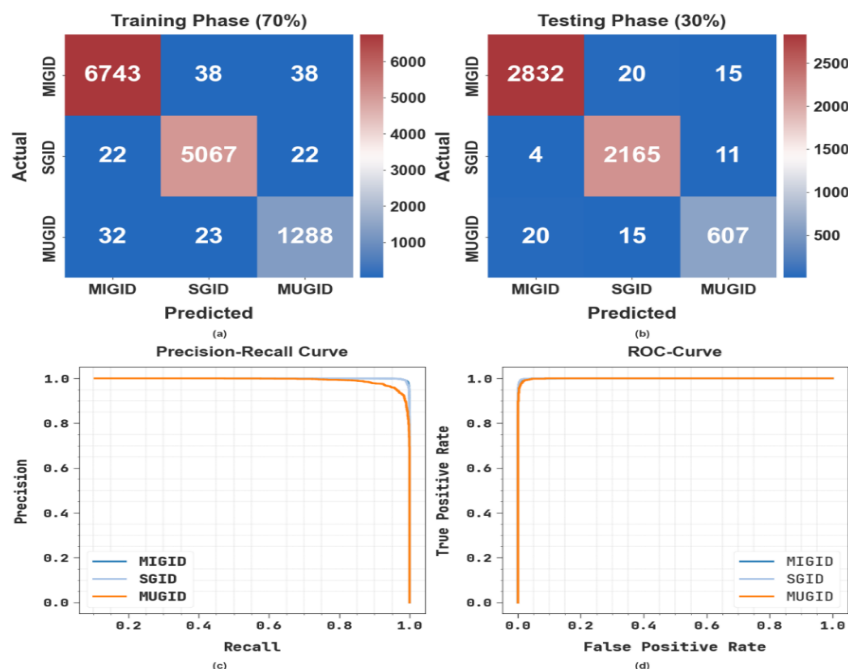


Fig. 3: Classifiers result of the XHTGT-AGDD approach with (a-b) Confusion matrices and (c-d) PR/ROC curves

Table 4 and Fig. 4 signify the genetic disease detection results of the XHTGT-AGDD approach on 70:30. The outcome indicated that the XHTGT-AGDD system accurately categorized the samples. On 70% TRAP, the XHTGT-AGDD method achieves average value with $accu_r$ of 99.12%, $preci_n$ of 97.85%, $recal_l$ of 97.98%, $F_{measure}$ of 97.92% and MCC of 97.21%, correspondingly. For the meantime, on 30% TESP, the XHTGT-AGDD system attains average value with $accu_r$ of 99.00%,

$preci_n$ of 97.82%, $recal_l$ of 97.55%, $F_{measure}$ of 97.68% and MCC of 96.90%, correspondingly.

Figure 5 displays the training (TRAINING) and validation (VALIDN) $Accu_r$ curves of the XHTGT-AGDD method. TRAINING $Accu_r$ constant rise and achieves a higher level, demonstrating efficient convergence and learning. The VALIDN $Accu_r$ closely followed the training tendency with a smaller and steady gap, signifying better generalizes and reduces overfitting.

Table 4: Genetic diseases detection result of the XHTGT-AGDD method with 70:30

1. Class Labels	2. $Accu_r$	3. $Preci_n$	4. $Recal_l$	5. $F_{measure}$	6. MCC
7. TRAP (70%)					
8. MIGID	9. 99.02	10. 99.21	11. 98.89	12. 99.05	13. 98.04
14. SGID	15. 99.21	16. 98.81	17. 99.14	18. 98.97	19. 98.33
20. MUGID	21. 99.13	22. 95.55	23. 95.90	24. 95.73	25. 95.24
26. Average	27. 99.12	28. 97.85	29. 97.98	30. 97.92	31. 97.21
32. TESP (30%)					
33. MIGID	34. 98.96	35. 99.16	36. 98.78	37. 98.97	38. 97.93
39. SGID	40. 99.12	41. 98.41	42. 99.31	43. 98.86	44. 98.15
45. MUGID	46. 98.93	47. 95.89	48. 94.55	49. 95.22	50. 94.61
51. Average	52. 99.00	53. 97.82	54. 97.55	55. 97.68	56. 96.90

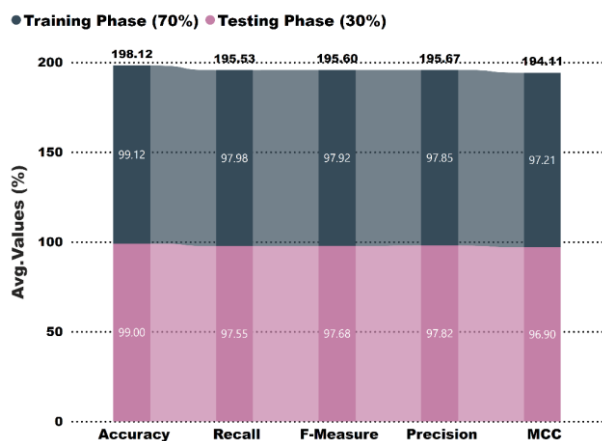


Fig. 4: Average value of the XHTGT-AGDD method with 70:30

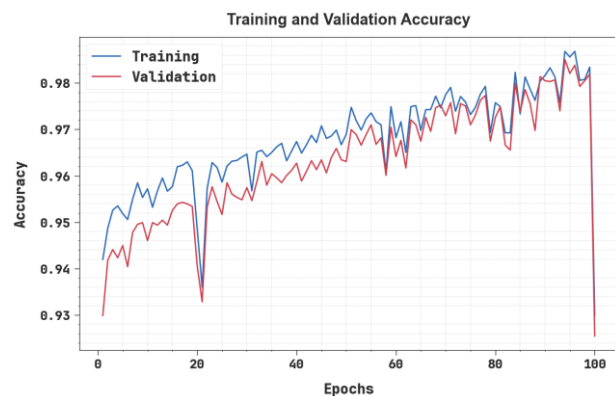


Fig. 5: $Accu_r$ curve of the XHTGT-AGDD technique

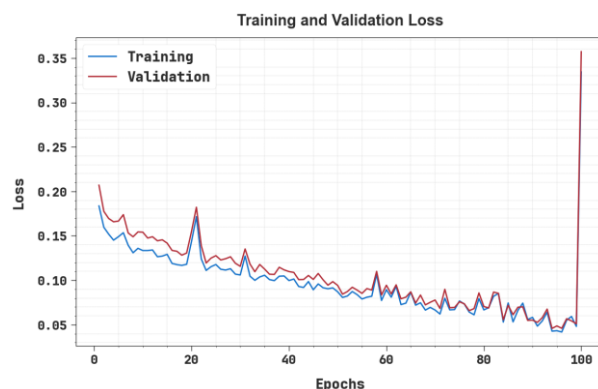


Fig. 6: Loss curve of the XHTGT-AGDD system

Figure 6 shows the TRAINING and VALIDN loss curves of the XHTGT-AGDD model. Both losses decline rapidly in the early epochs, implying fast learning, and then progressively reduced in a steady and smooth manner, reflecting efficient convergence. The near alignment between TRAINING and VALIDN loss, with only a lower constant gap, indicates lower overfitting and robust generalizes capacity. Since the proposed XHTGT-AGDD framework has high classification accuracy, it is necessary to assess the potential overfitting. The training and validation accuracy curves shown in Fig. 5 demonstrate similar behaviour during training and only a small difference is seen between the training and validation curves. Similarly, the learning curve and validation curve (shown in Fig. 6) indicate a stable convergence process with no major divergence.

The convergence characteristics of the proposed model can be seen from the training curve and validation curve. As the number of epochs increases, the accuracy of the training steadily rises, contributing to learning

discriminative patterns from genomic information. At the same time, the accuracy of the validation follows very closely the accuracy obtained on the training set, indicating good generalisation.

The superior performance of the XHTGT-AGDD method is highlighted through a brief comparison presented in Fig. 7. The result indicated that the RNN Method and Ensemble Learning-GWO model has exhibited the least performance, whereas CNN-MGP approach, SVM Classifier, LSTM-RNN model, DT

system, KNN algorithm, ANN classifier, AdaBoost system, and EfficientNet model have performed slightly the highest results. Moreover, the CNN-GRU and HDLMOA-DGD techniques have reported reasonable and closer results. At last, the proposed XHTGT-AGDD method has revealed maximal performance with $accuracy$ of 99.12%, $precis_n$ of 97.85%, $recal_i$ of 97.98%, and $F_{measure}$ of 97.92%.

Figure 8 represent the Computation Time (CT) of the XHTGT-AGDD methods with other methodologies.

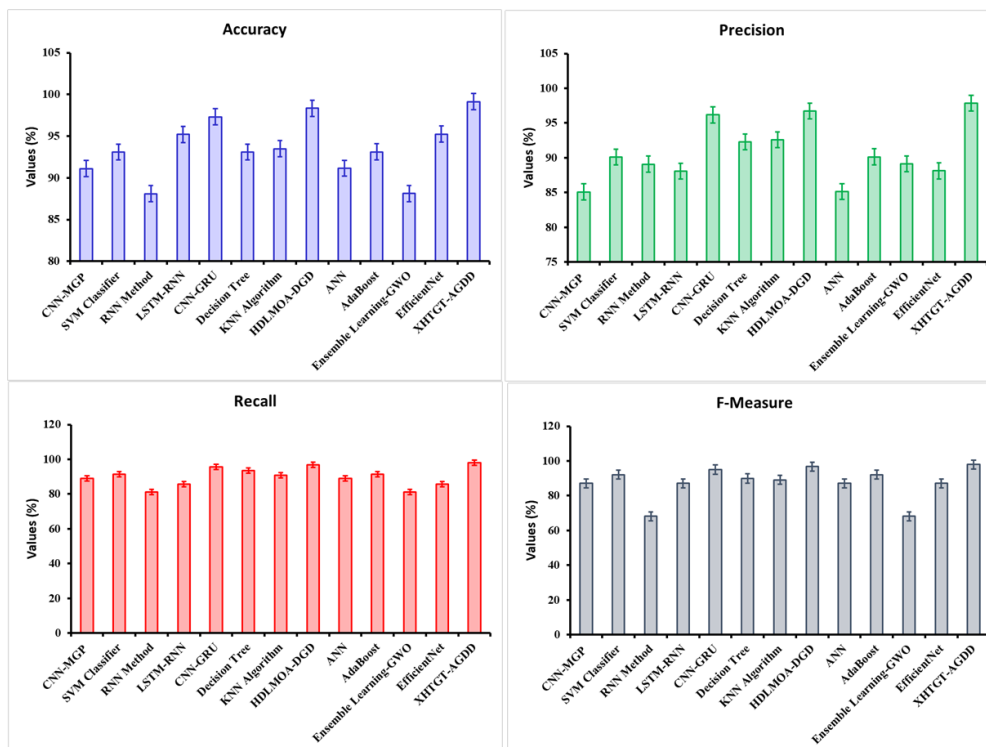


Fig. 7: Comparison result of the XHTGT-AGDD method with various measures

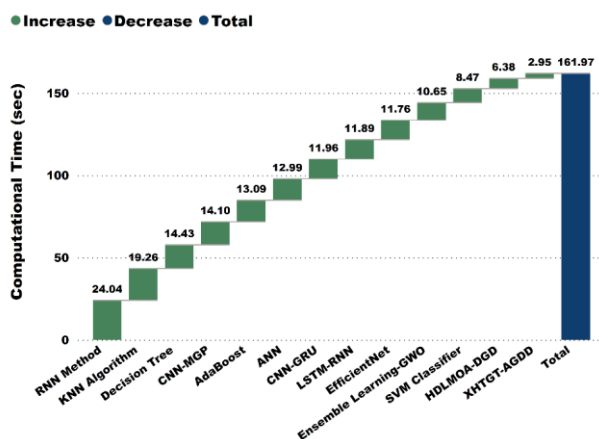


Fig. 8: Computation of the XHTGT-AGDD technique

Based on CT, the proposed XHTGT-AGDD model has achieved lesser time with CT value of 2.95s, while the other models such as CNN-MGP algorithm, SVM Classifier, RNN Method, LSTM-RNN system, CNN-GRU classifier, DT model, KNN algorithm, HDLMOA-DGD approach, ANN system, AdaBoost method, Ensemble Learning-GWO model, and EfficientNet technique have achieved higher time with CT of 14.10s, 8.47s, 24.04s, 11.89s, 11.96s, 14.43s, 19.26s, 6.38s, 12.99s, 13.09s, 10.65s, and 11.76s, correspondingly. The detailed results and discussion confirm that the XHTGT-AGDD model outperforms all other models.

Figure 9 depicts SHAP values to explain how various features influence the predictive methods for 3 classes. It emphasizes the most significant features impacting all classes, with “Disorder Subclass” constantly being the high contribution.

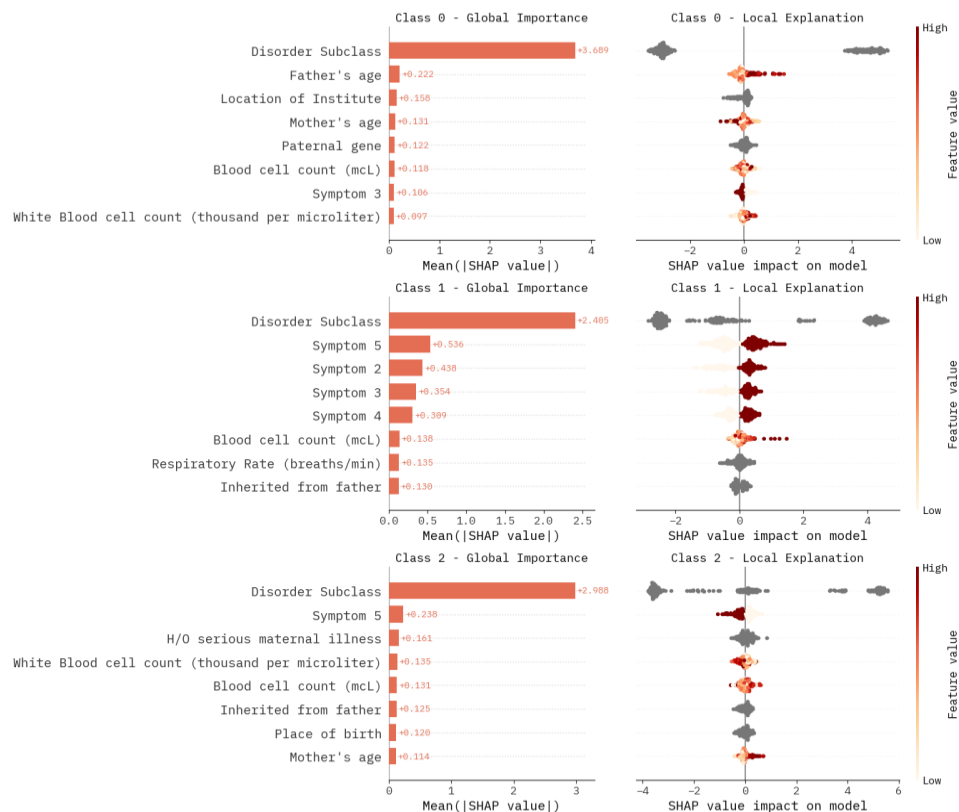


Fig. 9: SHAP based global feature importance and local explanation plots demonstrating the influence of features on the model predictions for class 0, class 1, and class 2

They display how differing feature values either decrease or increase the probability of a specific class prediction. Features include symptoms, blood cell counts, and inherited components also played important role but vary through classes. The visualization provides insight into that features the method based on most, helping interpretable and trust in its decisions.

Clinical Interpretation of SHAP Analysis

The SHAP analysis is useful to gain insight into factors that influence the model's predictions. Disorder Subclass has the greatest average contribution across all of the prediction categories, suggesting that this feature has the greatest discriminative power, being able to distinguish the different types of genetic disorder well. Other parental age attributes also have high contribution to prediction results, especially maternal and paternal age. The results are consistent with medical observations indicating that there are relationships between parental age and the presence of some inherited and chromosomal abnormalities.

Conclusion

In this manuscript, the novel XHTGT-AGDD methodology has been presented for improving the reliability and precision of genetic disorder detection from

genomic data by leveraging state-of-the-art methods. Primarily, quantile transformation is used to normalize the input data for effective analysis. Following preprocessing, feature selection was performed, and the most relevant feature subsets were identified from the genomic data using the FHO algorithm. The classification process utilized an HTGT architecture, which efficiently mapped intricate genomic relationships and their temporal dependencies. Moreover, the HO algorithm is implemented to refine model performance. Moreover, this study employed SHAP analysis to render the model transparent by quantifying the contribution of each genetic feature to the final prediction. The experimental result analysis of the XHTGT-AGDD model was conducted using a benchmark dataset. Improved performance of the XHTGT-AGDD model in terms of various metrics was reported by the simulation results with accuracy of 99.12%. Thus, reliable and effective performance for automatic genetic disorder detection from genomic data was provided by the proposed model. The proposed XHTGT-AGDD framework shows good performance, but there are some limitations. Firstly, the experimental evaluation was carried out with a public benchmark set and further testing with other independent, external sets could enhance the assessment of generalization. Second, transformer-based Graph

Architectures are more resource intensive than traditional machine learning models. Third, the current framework is based on structured, genomic and clinical features instead of sequencing data. Future research could include the expansion of the proposed framework to large scale genomic sequencing data sets, multi-center clinical trials, and clinical decision support systems.

Availability of Data and Materials

The data that support the findings of this study are openly available in [Predict the genetic disorders dataset-of genomes] at [https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes].

Acknowledgment

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, FEAT, Annamalai University, for providing the necessary research facilities and support to carry out this work.

Funding Information

The authors received no specific funding for this study.

Author's Contributions

Puvaneswari Subramaniapillai: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing, visualization.

Indirani Govindasamy: Conceptualization, validation, formal analysis, resources, writing review and edited, visualization, supervision.

Revathy Gurumurthy: Conceptualization, validation, resources, writing review and edited, visualization, supervision.

All authors reviewed and approved the final version of the manuscript.

Ethics

This study utilized publicly available genomic datasets and did not involve human participants or animal experimentation.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Alzahrani, A. A., & Alharithi, F. S. (2024). Machine learning approaches for advanced detection of rare genetic disorders in whole-genome sequencing. *Alexandria Engineering Journal*, 106, 582–593. <https://doi.org/10.1016/j.aej.2024.08.056>

Chen, J., Qi, J., Ao, Y., Wang, K., & Song, X. (2025a). A Solution Method for Non-Linear Underdetermined Equation Systems in Grounding Grid Corrosion Diagnosis Based on an Enhanced Hippopotamus Optimization Algorithm. *Biomimetics*, 10(7), 467. <https://doi.org/10.3390/biomimetics10070467>

Chen, Y.-M., Hsiao, T.-H., Lin, C.-H., & Fann, Y. C. (2025b). Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. *Journal of Biomedical Science*, 32(1), 16. <https://doi.org/10.1186/s12929-024-01110-w>

Gélin, M., Launay, É., & Vince, N. (2026). Whole-Exome Sequencing for the Identification of Genetic Factors Implicated in Severe Bacterial Infections: A Systematic Review. *The Journal of Infectious Diseases*, 233(3), e600–e610. <https://doi.org/10.1093/infdis/jiag006>

Guo, X., Feng, Y., Ji, X., Jia, N., Maimaiti, A., Lai, J., Wang, Z., Yang, S., & Hu, S. (2025). Shared genetic architecture and bidirectional clinical risks within the psycho-metabolic nexus. *EBioMedicine*, 111, 105530. <https://doi.org/10.1016/j.ebiom.2024.105530>

Gwak, S.-Y., Seo, J., Seo, G. H., Oh, J., Lee, H.-J., Kim, K., Cho, I., Shim, C. Y., Ha, J.-W., & Hong, G.-R. (2024). Role of Genetic Testing in Diagnosis and Prognosis Prediction in Hypertrophic Cardiomyopathy in Korea. *Journal of Korean Medical Science*, 39(50), e313. <https://doi.org/10.3346/jkms.2024.39.e313>

Jurgens, J. A., Barry, B. J., Chan, W.-M., MacKinnon, S., Whitman, M. C., & Ruiz, P. M. M. (2025). Expanding the genetics and phenotypes of ocular congenital cranial dysinnervation disorders. *Genetics in Medicine*, 27(4), 101216. <https://doi.org/10.1101/2024.03.22.24304594>

Kaggle. (2026). Predict the genetic disorders dataset of genomes. Kaggle Dataset Repository. *Kaggle Dataset Repository*. <https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes>

Keerthana, C. N., & Mahesh, R. (2025). Advancements in machine learning for predictive modelling of genetic disorders. *International Journal of Engineering Development and Research*, 13(4), 873–878.

Kirchhoff, A., Hustinx, A., Javanmardi, B., Hsieh, T.-C., Brand, F., Hellmann, F., Mertes, S., André, E., Moosa, S., Schultz, T., Solomon, B. D., & Krawitz, P. (2025). GestaltGAN: synthetic photorealistic portraits of individuals with rare genetic disorders. *European Journal of Human Genetics*, 33(3), 377–382. <https://doi.org/10.1038/s41431-025-01787-z>

Minten, T., Bick, S., Adelson, S., Gehlenborg, N., Amendola, L. M., & Boemer, F. (2025). Data-driven consideration of genetic disorders for global genomic newborn screening programs. *Genetics in Medicine*, 27(7), 101443. <https://doi.org/10.1101/2024.03.24.24304797>

- Moghaddam, H. K., Abtahizadeh, E., & Abolfathi, S. (2025). Sustainable water allocation under climate change: Deep learning approaches to predict drinking water shortages. *Journal of Environmental Management*, 385, 125600. <https://doi.org/10.1016/j.jenvman.2025.125600>
- Olivucci, G., Iovino, E., Innella, G., Turchetti, D., Pippucci, T., & Magini, P. (2024). Long read sequencing on its way to the routine diagnostics of genetic diseases. *Frontiers in Genetics*, 15, 1374860. <https://doi.org/10.3389/fgene.2024.1374860>
- Park, K. H., Lee, Y., Ding, W., Ryu, K. S., & Ryu, K. H. (2026). Transformer-based feature extraction approach for hematopoietic cancer subtype classification. *Computers in Biology and Medicine*, 203, 111466. <https://doi.org/10.1016/j.combiomed.2026.111466>
- Puvaneswar, S., & Indirani, G. (2026). A Deep Convolutional Autoencoder with Metaheuristic Optimization based Feature Reduction Framework for Genetic Disorder Detection Model. *Journal of Intelligent Systems and Internet of Things*, 18(1), 361–373. <https://doi.org/10.54216/jisiot.180128>
- Raj, S., Namdeo, V., Singh, P., & Srivastava, A. (2025). Identification and prioritization of disease candidate genes using biomedical named entity recognition and random forest classification. *Computers in Biology and Medicine*, 192, 110320. <https://doi.org/10.1016/j.combiomed.2025.110320>
- Raji, H., Dinesh, S., & Sharma, S. (2025). Inside the impulsive brain: a narrative review on the role of neurobiological, hormonal and genetic factors influencing impulsivity in psychiatric disorders. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, 61(1), 4. <https://doi.org/10.1186/s41983-024-00930-9>
- Shahid, A., Zahra, A., Aslam, S., Shamim, A., Ali, W. R., Aslam, B., Khan, S. H., & Arshad, M. I. (2026). Appraisal of CRISPR Technology as an Innovative Screening to Therapeutic Toolkit for Genetic Disorders. *Molecular Biotechnology*, 68(1), 71–94. <https://doi.org/10.1007/s12033-025-01374-z>
- Shahin, O. R., Alanazi, F. M., ElDeeb, M. K., Kuriri, F. A., Kwa, F. A. A., Alenazy, F. O., & Kamal, T. M. (2025). Construction of Saudi computational gene models: Applications in healthcare of prevalent genetic disorders. *Alexandria Engineering Journal*, 117, 13–26. <https://doi.org/10.1016/j.aej.2024.12.105>
- Srinivas, D. B., Deepika, S., Sindhu, G., Doddamani, N. R., & Deepika, K. M. (2025). A Machine Learning Approach to Classification and Prediction of Genetic Disorders. *2025 International Conference on Communication, Computer, and Information Technology (IC3IT)*, 1–6. <https://doi.org/10.1109/ic3it66137.2025.11341358>
- Wang, Y. (2025). HTGformer: Heterogeneous Temporal Graph Transformer. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2550–2554. <https://doi.org/10.1145/3726302.3730209>
- Xue, B., Niu, M., Sun, Y., Wang, L., Wu, C., Ding, Y., Wang, B., Peng, L., Li, X., Song, H., Yuan, W., Shi, W., Liu, J., Gao, C., Zhao, X., Zhang, Q., & Li, Z. (2026). Detection of pleiotropic genetic factors and critical brain cell types linking insomnia with psychiatric disorders. *SLEEPJ*, 49(1), zsaf317. <https://doi.org/10.1093/sleep/zsaf317>
- Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports*, 10(1), 15534. <https://doi.org/10.1038/s41598-020-72664-6>