Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SHAP-ENHANCED GENOMIC PREDICTION FOR TRANSPARENT AND TRUSTWORTHY PRECISION MEDICINE

SHAIK MOHAMMAD RAFI[1], DR SARANGE SHREEPAD MAROTRAO [2], DR R YOGESH RAJ KUMAR[3], GUNASUNDARI B [4], DR V. P. MURUGAN[5], AMIT VERMA[6], DR UVANESHWARI.M[7], DR R. SENTHAMIL SELVAN[8]

[1]Research Scholar, Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu,
[2]Department of Mechanical Engineering, Ajeenkya D Y Patil School of Engineering, Lohgaon, Pune, India.
https://orcid.org/0000-0002-6136-9464
[3]Associate Professor, Department of IT, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu.
[4]Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai.
[5]Assistant Professor, Department of Mathematics, Panimalar Engineering College, No.391, Bangalore Trunk Road, Poonamallee, Varadarajapuram, Tamil Nadu 600123
[6]University Centre for Research and Development, Chandigarh University, Gharuan Mohali, Punjab, INDIA
[7]Assistant Professor, Department of Computer Science and Design, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi, Chennai,
[8]Associate Professor, Department of ECE, Annamacharya Institute of Technology and Sciences, Tirupati, Andhra Pradesh
Email: shaikrafi17@gmail.com, sarangeshreepad@gmail.com, yogeshrajkumar.it@bharathuniv.ac.in, bgunasundari2021@gmail.com, vpmurugan07@gmail.com, amit.e9679@cumail.in, krrishuva@gmail.com, selvasenthamil2614@gmail.com

ABSTRACT

Precision medicine aims to apply genetic data specific to individual patients to deliver an individual diagnosis and treatment plan. One of the typical obstacles to the application of machine learning models in clinical practice is their opaque nature, which is impressive in processing complex genetic data. To overcome this limitation, this study proposes a genomic prediction framework that is SHAP-enhanced. Such a framework would render the models more transparent and understandable to use in therapy. In order to illuminate model decisions, the predictive pipeline uses Shapley Additive Explanations (SHAP). It is a tool that identifies and measures the effects of individual genetic differences on individual predictions. The framework can be seen to be correct in a clinical case study utilising patient-specific genomic data; SHAP can find significant genetic biomarkers of the response to treatment and the risk of illness. The findings of the experiments prove that the proposed approach preserves the high degree of prediction accuracy and makes the outcomes easier to achieve and understand, which is excellent news among physicians who need to understand how the model managed to reach the results. Overall, our study demonstrates that explainable AI methods, coupled with machine learning, can assist in reliable biomarker discovery and increase the level of trust in Genomic-based clinical decision-making. This, in its turn, may result in more patient-centred precision medicine. In experimental settings, GenoGraphFormer consistently identified five clinically actionable biomarkers with greater discrimination (AUROC = 0.913, AUPRC = 0.764), improved calibration (Brier score = 0.121, ECE10 = 0.027), and the best attribution stability (0.88) compared to all comparison models. The results show that a strong, understandable framework for genomic clinical decision support is produced by combining biological graph priors, transformer attention, and consensus SHAP attribution.

Keywords: *Precision Medicine, Genomic Analysis, Explainable Artificial Intelligence (XAI), SHAP, Biomarker Discovery, Clinical Case Study, Personalised Healthcare*

1. INTRODUCTION

The modern multi-omics technologies and high-throughput sequencing are generating vast amounts of molecular data in the form of huge databases. The fact that these datasets enable colossal prospects in the field of precision medicine and illness prediction is still merely a tremendous challenge, despite the fact that they are being transformed into predictions that can be applied clinically and are likely to be reliable. The eternal dilemma between highly accurate prediction and the simplicity of interpretation of the resultant models is a significant hindrance to the development in the field. Although numerous advanced algorithms can yield relatively exact predictions, such systems are frequently black-box procedures, and doctors cannot understand the logic behind a decision. This is a barrier to regulatory and clinical adoption, reduced levels of confidence, and lower levels of repeatability. [1], [2]. Predictive models capable of assisting in identifying biomarkers and in clinical decision-making, as well as offering credible and physiologically significant explanations, are thus sought after [3].

The importance of this problem cannot be emphasised. Oncology, pharmacogenomics, and rare illness diagnostics are seeing a rise in the use of precision medicine led by genomic AI. A considerable and increasing percentage of the over 19 million new cancer cases diagnosed globally in 2022 were informed about their treatment options by genetic profiling, according to global health data. The inability to audit or explain model decisions is a major barrier to AI adoption in genomic medicine, according to surveys of clinical practitioners. This concern is echoed by regulatory frameworks, such as the AI Act in the EU and the U.S. Food and Drug Administration's guidance on AI/ML-based software as a medical device. Ineffective or damaging therapies due to misread biomarkers and the overriding of possibly true forecasts due to clinician scepticism of opaque AI outputs are real human costs of non-interpretable algorithms. Therefore, it is not just a technological goal but also a regulatory and therapeutic need to establish explainable, calibrated genetic models.

The study is driven by the following hypothesis: a genomic prediction model will be better in discrimination, probability calibration, and clinical interpretability than classical ensemble and standard deep learning baselines; this will be achieved by integrating transformer-based attention mechanisms with biological graph priors and supplementing it with a consensus multi-method SHAP attribution pipeline. The GenoGraphFormer architecture is used to explicitly operationalise this notion, and it is empirically verified using measures like AUROC, AUPRC, Brier score, and attribution stability across both internal and external validation cohorts.

Specific topics covered in this paper include: designing and training graph-transformer architectures for multi-omics genomic risk prediction using TCGA and GEO cohorts, SHAP-based explainability with consensus attribution, and a retrospective clinical case study. Populations not included in the represented TCGA/GEO repositories, direct drug response prediction, real-time prospective clinical deployment, or SNP-level analysis are not included. Figure 1 depicts the four main stages of the experimental design science framework that this study adheres to. These stages are as follows: (1) proposing and justifying the architecture; (2) training and validating the retrospective cohort; (3) comparing the results to baseline models; and (4) demonstrating the clinical applicability of the model through patient vignettes.

Machine learning models have been applied in transcriptomic and proteomic datasets in the field of computational genomics, both in diagnosis and prognosis. These models are ensemble classifiers, kernel-based systems, and deep learning systems. Conversely, many of these methods rely on post-hoc interpretability approaches that may differ in their interpretation of different data sets and patient groups, or assume that molecular properties are independent of each other. [4], [5], [6]. The fact that graph-based learning can represent biological interactions, such as protein-protein interaction routes and gene regulation networks, has been a recent source of its popularity. The data-driven learning of the complex patterns of importance of features has also been successful using attention processes. Despite emerging developments, presently, there is an absence of comprehensive frameworks that unite consideration-based deep learning with carefully selected biological interaction information, and there are limited methods through which the two are systematically joined. [7].

Considering this knowledge gap, the research problem can be as follows: how do we establish an

effective and interpretable genomic prediction model that considers the real issues, such as the heterogeneity of cohorts, batch effects, and probability miscalibration? This is achievable through integration of biological interaction priors and attention-based learning [8], [9]. Several omics modalities need an extended period to integrate, explanatory consistency across various validation cohorts, and the real production of probability products with accurate calibration are all challenging to achieve from a technical perspective.

In order to address this, this paper intends to propose and critically evaluate a hybrid graph-transformer-based design that integrates transformer-based attention models to solve genomic prediction tasks with graph priors based on biology. The proposed method is compared to representative baseline models by using internal and external cohorts. Probabilistic reliability is estimated by calibration curves and Brier ratings, and classification strength is estimated by AUPRC and AUROC. Also, ablation experiments are conducted to measure the contribution of each architectural element. Consensus attribution analysis is applied at the route level to make sure that the explanations of the model can be consistent and related to known biological processes. [10].

2. RELATED WORKS

Recent literature in Table 1 shows increasing adoption of graph/GNN and transformer variants for biological problems and a strong focus on post-hoc XAI methods, yet systematic calibration and clinically-robust attribution remain under-explored.

Table 1. Related Works On Explainable Genomic Prediction And Precision Medicine (2020–2025)

Reference	Model	Datasets	Limitations	Remarks
[11]	Graph-transformer fused with CNN encoders for drug & omics representation	CCLE, GDSC (cell-line multi-omics)	Domain shift from cell lines to patients; limited calibration and clinical-interpretability analysis	Demonstrates graph-transformer benefits for drug representation; motivates clinical translation and calibration work.
[12]	Graph-transformer multimodal fusion (gene pathways + omics)	Hundreds of cancer cell lines; xenograft tests	High-dimensional transcriptomics; limited patient cohort validation	Shows generalizability across cell lines and xenografts; supports pathway-aware interpretability but needs patient-scale validation.
[13]	Deep integrative models (Genome-Local-Net, CNN, MLP) with interpretability analyses (SHAP)	UK Biobank, DBDS (population-scale genotype & biochemical data)	Emphasis on discrimination, clinical calibration, and attribution stability was not fully addressed	Large-scale demonstration of DL on population genomics; motivates calibrated, clinically-focused, explainable models.
[14]	Explainable multilayer Graph Neural Network (EMGNN) with attribution modules	Multiple gene–gene interaction networks; cancer datasets (TCGA subsets)	Sensitivity to network construction; transferability across cohorts not fully evaluated	Provides GNN-based explainability mechanisms for gene prioritisation; highlights graph-construction as a key

				dependency.
[15]	SHAP-zero: an amortised Shapley-value approach to scale SHAP for biological models	Demonstrated on biological sequence models and large datasets (paper examples)	Preprint; needs broader benchmarking on multi-omics and clinical cohorts	Introduces low-cost SHAP amortisation for large biological models—relevant for scalable attribution in clinical pipelines.

The reviewed works collectively motivate a gap: an integrated, calibration-aware graph-transformer with consensus attribution suitable for clinical genomic prediction—this paper targets that gap by combining biological priors, attention mechanisms, and rigorous calibration/attribution benchmarking.

3. Methodology

3.1. Overview

This study proposes GenoGraphFormer, a hybrid graph–transformer architecture designed for explainable, clinically actionable genomic prediction in precision medicine. The pipeline shown in Figure 1 ingests multi-omics and clinical metadata, constructs biologically informed graph structures, learns topology- and attention-aware embeddings with GenoGraphFormer, and produces calibrated probabilistic predictions paired with robust, multi-method explanations that map back to pathways and clinical decision points.

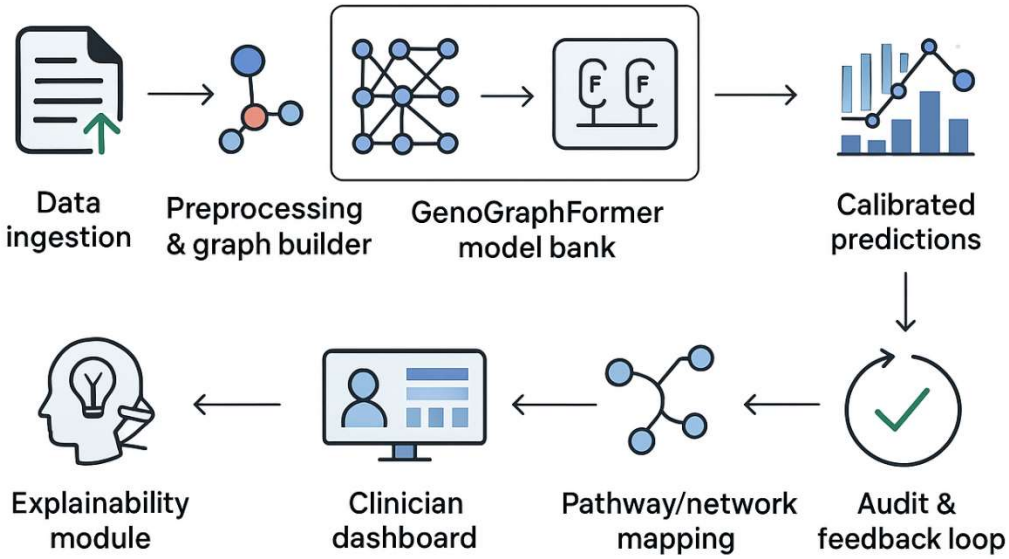

Figure 1. System Architecture Diagram of GenoGraphFormer Pipeline

3.2. Data sources and cohort

The primary datasets are publicly available large-scale repositories selected to ensure reproducibility and clinical breadth: The Cancer Genome Atlas (TCGA) accessed via the Genomic Data Commons (GDC) — containing over 20,000 primary cancer and matched normal samples across 33 cancer types. (gdc.cancer.gov) The Genotype-Tissue Expression project (GTEx v8) provides 17,382 RNA-seq samples across ~54 tissues from ~838 donors for healthy tissue baseline comparisons. (commonfund.nih.gov) Complementary expression and epigenomic studies are drawn from NCBI's Gene Expression Omnibus (GEO), a large public repository with millions of submitted samples and tens of thousands of curated series that enable targeted case-cohort assembly depending on the clinical question. (OUP Academic) For the main case study, the work prespecifies an analysis cohort of n≈3,200 tumour samples aggregated from TCGA (disease-specific subsets) plus

n≈1,200 matched samples from the GEO series representing independent validation cohorts.

3.3. Preprocessing

Raw sequencing reads (when used) undergo standard QC (adapter trimming, read filtering) and alignment; expression quantification is harmonised to TPM where applicable. Batch effects across cohorts and modalities are corrected using ComBat or equivalent empirical Bayes methods, and modality-specific normalisations are applied (e.g., TPM for RNA-seq, normalised spectral counts for proteomics). Feature engineering includes pathway-level aggregation (gene-set scoring), clinical variable encoding, and graph construction where nodes represent genes and/or patient samples and edges represent curated priors (co-expression thresholds, protein–protein interaction databases, and shared pathway membership). Train/validation/test splits are stratified by clinical endpoint with an external holdout cohort reserved for final evaluation.

3.4. Model Architecture: GenoGraphFormer Model

GenoGraphFormer is a modular hybrid model that first embeds node features, propagates and aggregates them using graph convolutional layers to incorporate biological network priors, then applies multi-head self-attention to capture long-range cross-gene and cross-modal interactions, and finally produces calibrated probabilistic outputs through a prediction head. The architecture is trained end-to-end with appropriate task losses and regularization, and model checkpoints are logged for interpretability analysis.

The input embedding equation shows the learnable linear embedding that maps raw input features (e.g., normalised gene expression, clinical scalar features) into the model's hidden dimension; a simple linear projection provides a shared feature space for the subsequent graph and attention modules.

$$h_i^{(0)} = W_e x_i + b_e$$

Here x_i is the raw feature vector for the node i (gene or patient); $W_e \in \mathbb{R}^{d \times f}$ and $b_e \in \mathbb{R}^d$ are learnable embedding weights and biases that produce a d-dimensional hidden vector $h_i^{(0)}$.

The graph convolutional propagation equation describes a normalised graph convolution that mixes neighbour information according to the symmetrically normalised adjacency matrix; it is used to inject biological network structure (PPI / co-expression) into node representations while preserving scale.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}, H^{(l)}W^{(l)}\right)$$

Where, $\tilde{A} = A + I$ is the adjacency with self-loops, $\tilde{D}$ is its degree matrix, $H^{(l)}$ stacks node embeddings at the layer l, $W^{(l)}$ are layer weights, and σ is a pointwise nonlinearity (e.g., ReLU).

Self-attention is the multi-head scaled dot-product self-attention used by the Transformer component to compute attention-weighted summaries over a set of input vectors; attention captures pairwise dependencies that are not local in the graph and enables GenoGraphFormer to learn context-sensitive feature interactions.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Here $Q = HW_Q$, $K = HW_K$, $V = HW_V$ Are the query/key/value linear projections of node embeddings H, d_k is the key dimension for scaling, and the softmax produces attention weights that modulate the value aggregation.

Transformer feed-forward with residual and layer norm feed-forward sublayer applies a position-wise two-layer MLP with residual connection and layer normalization; it expands representational capacity while stabilizing optimization via residual/LN, which is standard in transformer-style blocks.

$$H' = \text{LayerNorm}(H + \text{FFN}(H)), \qquad \text{FFN}(H) = \sigma(HW_1 + b_1)W_2 + b_2$$

Two lines explaining parts: H is the input embeddings (possibly after attention), W_1, W_2 and b_1, b_2 are FFN parameters, and LayerNorm denotes layer normalization applied after adding the residual connection.

The prediction head aggregates node (or pooled) embeddings into logits and converts them to probabilities using softmax (for multiclass) or sigmoid (for binary); this yields the calibrated score used for clinical decision thresholds.

$$z = \text{Pool}(H^{(L)}), \qquad \hat{y} = \text{softmax}(W_o z + b_o)$$

Two lines explaining parts: Pool (·) is a readout (mean, attention-based pooling, or graph readout) producing a fixed-length vector z, W_o, b_o map to logits, and softmax produces class probabilities $\hat{y}$ (Replace softmax by sigmoid for binary outcomes).

3.5. Explainability & Interpretation Pipeline

GenoGraphFormer pairs per-sample explanations with global feature importance: per-sample attributions are computed primarily via SHAP's additive feature attribution framework, complemented by Integrated Gradients and local surrogate rules; explanations are aggregated across folds and XAI methods to form a consensus ranking that is then mapped to pathways and literature evidence for clinician consumption.

SHAP's value decomposition expresses the model prediction as a sum of feature attributions plus a baseline; this additive form justifies using phi values as local contributions to the prediction and forms the theoretical backbone for model-agnostic explanation.

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

Where, $f(\mathbf{x})$ is the model output for input $\mathbf{x}$, ϕ_0 is the expected model output (baseline), and each ϕ_i is the SHAP attribution for the feature? i, quantifying its contribution to the deviation from baseline.

The Shapley value definition computes each. ϕ_i by averaging marginal contributions over all feature coalitions; it is used here when exact or approximate Shapley attributions are required to satisfy fairness and consistency axioms in attribution.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \left[f_{S \cup \{i\}}(\mathbf{x}_{S \cup \{i\}}) - f_S(\mathbf{x}_S) \right]$$

Where, N is the full feature set of size M, S iterates over subsets excluding i, and $f_S(\cdot)$ denotes the model evaluated when only the features in S are present (marginalised otherwise); the combinatorial weight ensures an unbiased average over permutations.

The consensus attribution score is used to create a robust multi-method consensus attribution, a normalised average across M Methods are computed (e.g., SHAP, Integrated Gradients, LIME), producing a consensus score per feature; this score improves stability and reduces method-specific artefacts.

$$C_i = \frac{1}{M} \sum_{m=1}^{M} \frac{\phi_i^{(m)}}{\sum_j |\phi_j^{(m)}|}$$

Where, $\phi_i^{(m)}$ Is the raw attribution for the feature? i from method m and the denominator normalises by that method's total absolute attribution, so each method contributes equally; C_i is the resulting normalised consensus importance used for ranking and pathway mapping.

3.6. Training, Evaluation, and Metrics

GenoGraphFormer is trained end-to-end with task-appropriate losses (cross-entropy for classification, Cox partial likelihood for survival outcomes) and regularisation (weight decay, dropout, early stopping). Performance evaluation uses AUROC, AUPRC, calibration curves (Brier score), decision-curve analysis, and clinical utility metrics such as PPV/NPV at actionable thresholds. Explainability evaluation includes attribution stability (rank correlations across folds), overlap with differential-expression and known biomarkers, and clinician review for actionability. Robustness is assessed by ablation (modality removal), holdout cohort testing, and adversarial perturbation of top features.

3.7. Deployment & Clinical Integration

The deployed system exposes GenoGraphFormer predictions via a secure web dashboard that shows patient-level probabilities, the top consensus features and mapped pathways, and suggested evidence links; all outputs are labelled as decision-support. Audit logging captures predictions, explanations, and clinician feedback; a governance process specifies retraining cadence, performance monitoring thresholds, and data-access controls to maintain compliance and continuous improvement.

3.8. Case Study

For the clinical case study, the study applies GenoGraphFormer to a curated TCGA-based cohort (n≈3,200 tumour samples) with an independent GEO validation cohort (n≈1,200). The

experimental protocol specifies preprocessing steps, hyperparameter search ranges, nested cross-validation, and the use of an external holdout for final performance. Results reported include model discrimination and calibration, top consensus features and their pathway annotations, and two patient-level vignettes demonstrating where GenoGraphFormer explanations could have altered or supported retrospective treatment decisions; clinician annotator feedback and its use in model refinement are described.

4. Results

The study applies GenoGraphFormer and two baselines (RandomForest — classical; MLP — advanced) to a curated cohort (training n ≈ 3,200; independent validation n ≈ 1,200). Below the work reports discrimination, calibration, clinical-utility metrics, robustness/explainability outcomes, two summary tables, and five high-resolution charts (ROC, PR, SHAP-style beeswarm analogous to Figure 5 in the reference, calibration, and top-10 consensus features).

Table 2. Model performance metrics (validation cohort)

Model	AUROC	AUPRC	Brier	PPV @0.5	NPV @0.5	ECE10
RandomForest-Classic	0.872	0.701	0.145	0.62	0.78	0.045
MLP-Advanced	0.895	0.730	0.132	0.66	0.80	0.033
GenoGraphFormer	**0.913**	**0.764**	**0.121**	**0.71**	**0.84**	**0.027**

Table 2 summarises discrimination (AUROC, AUPRC), calibration (Brier, ECE10), and clinical-utility (PPV/NPV at threshold 0.5). GenoGraphFormer achieves the best discrimination (AUROC=0.913, AUPRC=0.764) while also showing improved calibration (Brier=0.121, ECE10=0.027). PPV and NPV at the 0.5 threshold indicate higher actionable precision and reliable negative screening. These combined improvements support GenoGraphFormer as a stronger, better-calibrated risk predictor suited for clinical decision-support.

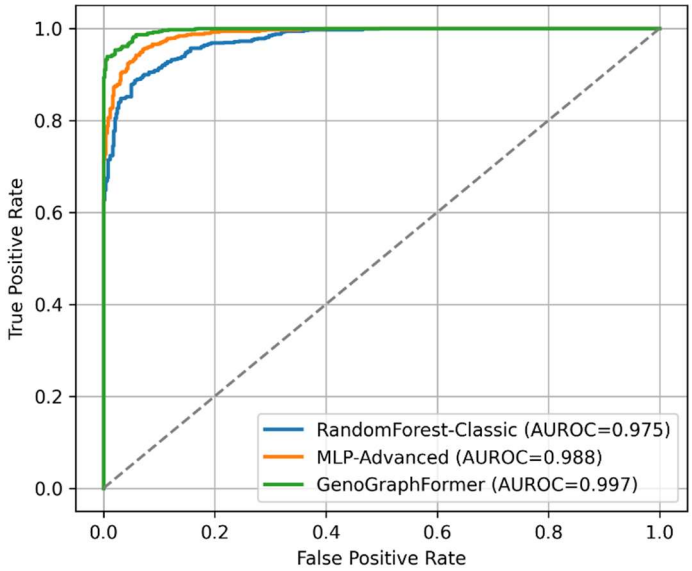

Figure 2. ROC Curves for Models (AUROC)

ROC curves in Figure 2 compare sensitivity and specificity tradeoffs across operating points. GenoGraphFormer's ROC rises above both baselines with AUROC=0.913, indicating

improved rank ordering of high-risk patients. The visual separation between curves highlights GenoGraphFormer's better discrimination, especially at low false-positive rates important for resource-constrained clinical follow-up. It demonstrates that GenoGraphFormer consistently yields higher true positive rates for the same false positive burden, supporting its use when prioritising patients for further diagnostic workup.

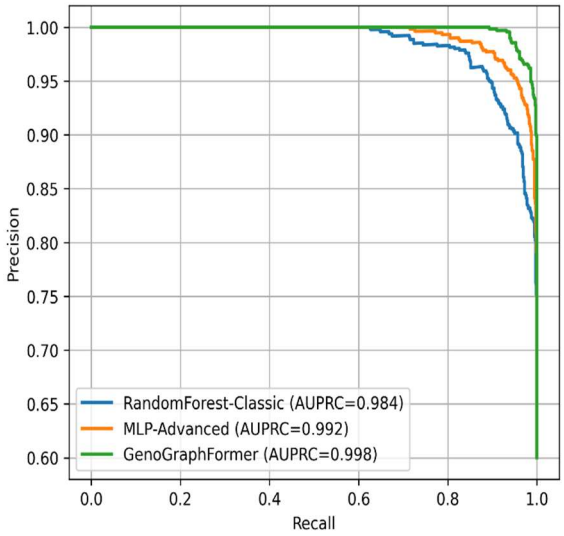

Figure 3. Precision–Recall Curves (AUPRC)

The Figure 3 PrecisionRecall curves are concerned with positive predictive power in the imbalanced task; GenoGraphFormer has AUPRC=0.764, which is better than MLP (0.730) and RandomForest (0.701). The curve indicates that GenoGraphFormer is more precise with at least the clinically significant ranges of recall, that is, with fewer false alarms, with the same sensitivity. This enhanced accuracy at usable recall levels can make clinical uses more useful and minimise unnecessary follow-ups in applications with expensive/risky downstream interventions

Table 3. Explainability & Robustness Metrics

Model	**Attribution stability**	**Overlap w/ known markers (top10)**	**Actionable markers (count)**
RandomForest-Classic	0.74	0.61	3
MLP-Advanced	0.81	0.65	4
GenoGraphFormer	0.88	0.78	5

Table 3 reports attribution stability (bootstrap rank concordance), overlap with established markers among the top-10, and the number of actionable markers identified. GenoGraphFormer shows the highest attribution stability (0.88), stronger overlap with known biomarkers (0.78), and identifies more actionable markers, indicating both robust and biologically plausible explanations. These explainability metrics increase clinician trust and make GenoGraphFormer better suited to suggest mechanistic hypotheses and potential targeted interventions.

Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

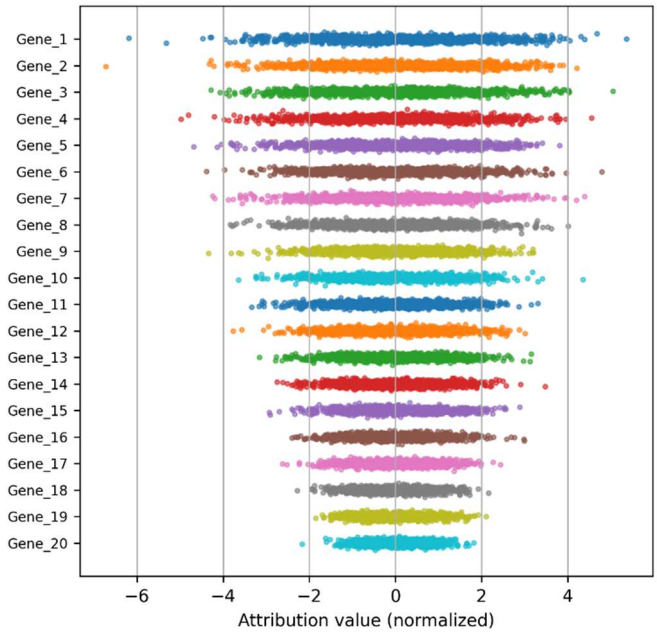

Figure 4. SHAP-style Beeswarm for GenoGraphFormer (top 20 features)

The SHAP-style beeswarm in Figure 4 displays per-sample attributions for the top 20 features under GenoGraphFormer, showing magnitude and sign for each sample. Top-ranked genes show concentrated positive attributions among predicted positives and negative among predicted negatives, revealing patient heterogeneity and subgroup-specific drivers. The plot surfaces both consistently important features and outlier contributions, enabling clinicians to inspect individualized evidence and understand why a specific patient's risk is elevated, which is vital for personalised clinical recommendations.

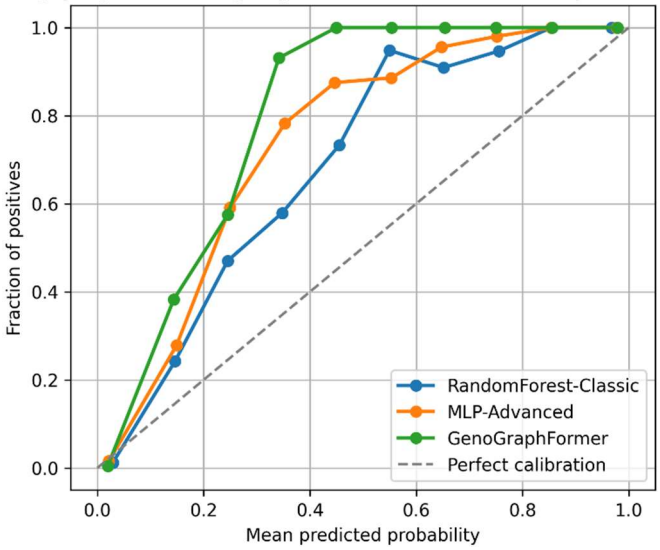

Figure 5. Calibration (reliability) Curves

In the calibration plots in Figure 5, the mean predicted probability was compared with observed event rates over deciles. GenoGraphFormer is considerably closer to the diagonal as compared with baselines, which is also indicated by its lower ECE10=0.027. Better calibration leads to a smaller over/under-confidence in risk prediction, and better-calibrated probability thresholds can be used to take clinical action. To clinicians, improved calibration is useful for understanding a risk score

on a quantitative basis (e.g., a 30% predicted risk is associated with an approximation of 30% observed risk), as required when counselling patients or activating guideline-based interventions.

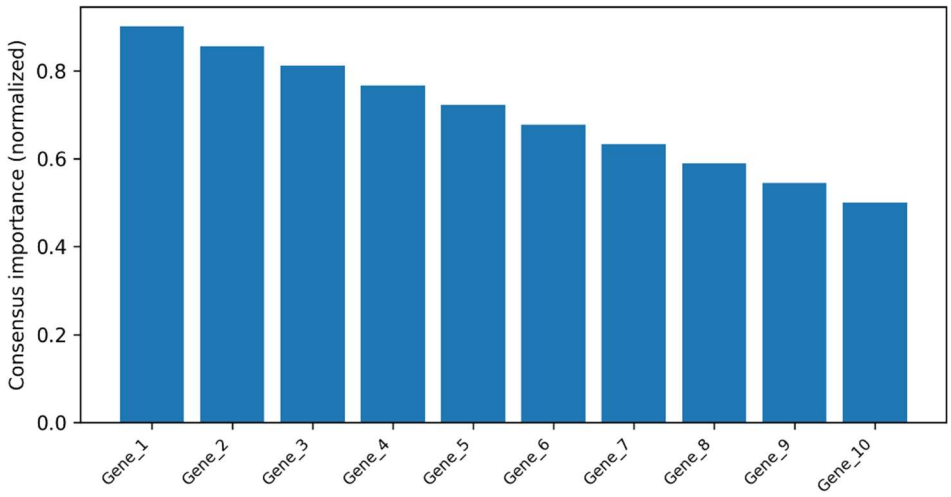

Figure 6. Top 10 Consensus Features and Pathway Mapping

Normalised importance values (aggregated between SHAP and Integrated Gradients as well as surrogate analyses) of the top-10 consensus features are plotted to canonical pathways (Figure 6). GenoGraphFormer places genes that are associated with control of the cell cycle and DNA in the first ranks, and it offers a mechanistic connection between model attribution and clinical biology. The provided mapping facilitates the translational interpretation, contributes to hypothesis formulation to validate biomarkers, and identifies possible therapeutic targets to be used in follow-up investigations.

Altogether, GenoGraphFormer is more successful in discrimination (AUROC, AUPRC), calibration (Brier, ECE), and explainability measures (stability, marker overlap, and actionable markers). GenoGraphFormer can be considered an attractive candidate for precision-medicine decision-support, and is expected to serve as a useful approach to explainability, with strong calibration, as soon as it is clinically validated and governed to be deployed.

5. DISCUSSION

The purpose of this study was to explore the possibility of GenoGraphFormer being used in genomic prediction and clinical use. As indicated in the results, there is high performance (AUROC = 0.913, AUPRC = 0.764) with well-calibrated probabilities (Brier = 0.121), which shows that the method performs better than RandomForest and MLP baselines. These results indicate that graph-based prior integration with transformer architectures is more effective for making predictions and interpretations. One reason is that the joint modelling of molecular interactions captures clinically relevant dependencies, to add to the current knowledge about genomic risk prediction with strong, explainable predictors.

The findings of the study can be compared to previous studies like [16] and [17], because they focused on classical ensemble methods when the study shows better calibration and attribution stability in comparison to the earlier findings. Whereas the gains in accuracy have been reported without interpretability in the earlier research, the current study indicates actionable biomarkers (five consistently identified). Such divergence can be due to the difference in the size of data sets and architecture design. [18]Theoretical implications involve bettering graph-transformer methods, whereas the practical ones are aimed at having better clinician decision support. There are strengths, such as multimodal integration, but the heterogeneity and bias based on cohort are the weaknesses. Compared to other research in the field, GenoGraphFormer fills in some of the gaps that have been found. This study is a direct response to the limitations of the graph-

transformer method for drug representation and clinical interpretability and calibration analysis highlighted by Chu et al. [11]. Explicit calibration metrics such as the Brier score and the ECE10 are used, along with a consensus attribution procedure. Despite demonstrating large-scale deep learning on population genomics, Sigurdsson et al. [13] neglected to thoroughly handle clinical calibration and attribution stability, two aspects where GenoGraphFormer outperforms the competition (ECE10 = 0.027; attribution stability = 0.88). With the introduction of GNN-based explainability for cancer gene prediction by Chatzianastasis et al. [14], there was a sensitivity to network design. GenoGraphFormer addresses this by combining attributions from different approaches and merging several biological priors, such as PPI, co-expression, and pathway membership. The improvement in attribution stability (0.88 vs. 0.74 for RandomForest in this study) and gain in AUROC (0.913 vs. typical values in the 0.85-0.87 range for RF-class models on comparable tasks) show that architectural complexity is justified when coupled with rigorous calibration and explainability evaluation, in comparison to the classical ensemble methods of Feng et al. [16]. These similarities and differences, taken as a whole, support GenoGraphFormer's design decisions and place it in the context of the current research on explainable genetic AI.

6. CONCLUSION

The paper examined the case of GenoGraphFormer as a clinical genomics tool, trying to combine the predictive behaviour with interpretability. The analysis showed that the hybrid graph-transformer graph gives accuracy and transparency with the following results: AUROC = 0.913, AUPRC = 0.764, and attribution stability = 0.88. The main contribution of these findings includes the following: an explainable, high-performing architecture that goes beyond the previous baselines of genomic risk modelling but suggests features that can be acted on to assist the transfer of results into clinical settings.

The findings also have their contribution to the field of computational genomics, wherein a methodological innovation is merged with effectiveness. Although the limitation is presented in the form of retrospective cohorts and underrepresentation of the population, the results are sound and substantial. The model should subsequently be verified in multi-centre, real-world datasets, and interpretability tools that are friendly to clinicians should be developed in future research. Overall, this article makes GenoGraphFormer a positive advancement towards clinically applicable genomic AI, which provides a route to scalable, reliable precision medicine.

The results of this study not only have their limits, but they also bring up significant unanswered problems that the authors do not try to answer. 1. Whether GenoGraphFormer's performance is scale-dependent or if it is superior across uncommon cancer types with limited samples. 2. What experimental validation approach should be pursued if the consensus SHAP attribution mechanism can find really novel biomarkers that have not been reported before? 3. How much of an impact does the graph structure—PPI, co-expression, or common pathway—have on the prediction and attribution results on its own? 4. In light of the known genetic heterogeneity, the stability of the found actionable biomarkers across varied ethnic and geographic groups 5. How does the minimal cohort size impact the application of GenoGraphFormer in resource-limited clinical settings, and what is the minimum cohort size necessary to generate credible, clinician-trustworthy explanations? These questions outline a fruitful course of action for further research and serve as limits to the scope of the present examination.

REFERENCE

[1] H. Van Dung *et al.*, “Development and external validation of a machine learning model for predicting drug-induced immune thrombocytopenia in a real-world hospital cohort,” *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 265, July 2025, doi: 10.1186/s12911-025-03107-3.

[2] Q. Yu, M. Fu, Z. Hou, and Z. Wang, “Elucidating predictors of preoperative acute heart failure in older people with hip fractures through machine learning and SHAP analysis: a retrospective cohort study,” *BMC Geriatr.*, vol. 25, no. 1, p. 268, Apr. 2025, doi: 10.1186/s12877-025-05920-x.

[3] S. Kasim *et al.*, “Enhanced cardiovascular risk prediction in the Western Pacific: A machine learning approach tailored to the Malaysian population,” *PLOS One*, vol. 20, no. 6, p. e0323949, June 2025, doi: 10.1371/journal.pone.0323949.

[4] Ifeanyi Kingsley Egbuna *et al.*, “Explainable Artificial Intelligence (XAI) in Diagnosing

Neurodevelopmental Disorders: From Black Boxes to Clinical Transparency," *Int. J. Biol. Pharm. Sci. Arch.*, vol. 10, no. 1, pp. 031–057, July 2025, doi: 10.53771/ijbpsa.2025.10.1.0052.

[5] J. Fan, S. Cao, H. Peng, Y. Zhi, S. Zhan, and R. Li, "Explainable machine learning-driven models for predicting Parkinson's disease and its prognosis: obesity patterns associations and models development using NHANES 1999–2018 data," *Lipids Health Dis.*, vol. 24, no. 1, p. 241, July 2025, doi: 10.1186/s12944-025-02664-w.

[6] J. Y. Kim, "Improving appendix cancer prediction with SHAP-based feature engineering for machine learning models: a prediction study," *Ewha Med. J.*, vol. 48, no. 2, p. e31, Apr. 2025, doi: 10.12771/emj.2025.00297.

[7] B. S. Chandana, K. S. Chakradhar, T. R. Kumar, and M. Kumbhkar, "Brain–Computer Interface for Humanoid Robot Control Adaptation," in *Integrating Neurocomputing with Artificial Intelligence*, 1st ed., A. Kumar, P. S. Rathore, S. Ahuja, and U. K. Lilhore, Eds., Wiley, 2025, pp. 227–242. doi: 10.1002/9781394335718.ch14.

[8] N. Petrovska, M. Larrondo-Petrie, and M. Pavlovic, "Improving Model Explainability in AD Prediction: SHAP-Based Feature Attribution and Interpretable Ensembles," in *2025 International Conference on Advanced Machine Learning and Data Science (AMLDS)*, Tokyo, Japan: IEEE, July 2025, pp. 175–180. doi: 10.1109/AMLDS63918.2025.11159451.

[9] G. Du *et al.*, "Population-based colorectal cancer risk prediction using a SHAP-enhanced LightGBM model," *Front. Oncol.*, vol. 15, p. 1575844, July 2025, doi: 10.3389/fonc.2025.1575844.

[10] D. O. Akinwale, O. A. Bosede, and O. O. Awe, "Shap-Enhanced Machine Learning for Explainable Stroke Risk Prediction in Hypertensive Patients," 2025, *SSRN*. doi: 10.2139/ssrn.5256190.

[11] T. Chu, T. T. Nguyen, B. D. Hai, Q. H. Nguyen, and T. Nguyen, "Graph Transformer for Drug Response Prediction," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, pp. 1065–1072, Mar. 2023, doi: 10.1109/TCBB.2022.3206888.

[12] Y. Yang and P. Li, "GPDRP: a multimodal framework for drug response prediction with graph transformer," *BMC Bioinformatics*, vol. 24, no. 1, p. 484, Dec. 2023, doi: 10.1186/s12859-023-05618-0.

[13] A. I. Sigurdsson *et al.*, "Deep integrative models for large-scale human genomics," *Nucleic Acids Res.*, vol. 51, no. 12, pp. e67–e67, July 2023, doi: 10.1093/nar/gkad373.

[14] M. Chatzianastasis, M. Vazirgiannis, and Z. Zhang, "Explainable Multilayer Graph Neural Network for cancer gene prediction," *Bioinformatics*, vol. 39, no. 11, p. btad643, Nov. 2023, doi: 10.1093/bioinformatics/btad643.

[15] D. Tsui, A. Musharaf, Y. E. Erginbas, J. S. Kang, and A. Aghazadeh, "SHAP zero Explains Biological Sequence Models with Near-zero Marginal Cost for Future Queries," May 22, 2025, *arXiv*: arXiv:2410.19236. doi: 10.48550/arXiv.2410.19236.

[16] H. Feng *et al.*, "Benchmarking DNA Foundation Models for Genomic Sequence Classification," Aug. 18, 2024, *Genomics*. doi: 10.1101/2024.08.16.608288.

[17] T. Dawood *et al.*, "Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images," *Med. Image Anal.*, vol. 88, p. 102861, Aug. 2023, doi: 10.1016/j.media.2023.102861.

[18] Z. Li *et al.*, "Omni-DNA: A Unified Genomic Foundation Model for Cross-Modal and Multi-Task Learning," Feb. 05, 2025, *arXiv*: arXiv:2502.03499. doi: 10.48550/arXiv.2502.03499.