Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1192

Explainable Multimodal Fusion of Genomic and Clinical Data for Multi-Disease Prediction: A Deep Learning Approach

C. Raghavendra[1*], R. Suneetha Rani[2], CH. N. Santhosh Kumar[3], Veeramachaneni Dhanasree[4], D. Swapna[5], and Goguri Rashmitha[5]

[1]Department of CSE (Cyber Security), CVR College of Engineering, Hyderabad, India.
[2]Department of CSE, BVRIT HYDERABAD College of Engineering for Women. Hyderabad, India.
[3]CSE Department, ANURAG ENGINEERING COLLEGE, KODADA, India.
[4]Department of CSE (CS), Geethanjali Collage of Engineering and Technology, Hyderabad, India.
[5]Department of CSE, BVRIT HYDERABAD College of Engineering for Women. Hyderabad, India.
[*]Corresponding Author: C. Raghavendra. Email: crg.svch@gmail.com

Received: November 04, 2025 **Accepted:** January 19, 2026

Abstract: Precision medicine is an effort to customize healthcare treatment based on individual-specific genetic, clinical, and environmental traits. This study introduces an explainable AI platform to fuse genomic and clinical information to enhance disease prediction and individualized treatment regimens. Pre-processed, normalized, and multi-modal datasets consisting of whole-genome sequencing, gene expression data, and electronic health records were integrated through a hybrid data fusion process. Feature engineering and dimensionality reduction techniques were utilized to discover biological and clinical significant patterns, which was followed by meticulous training of a multi-layer neural network for prediction. Explainability was also coupled with SHAP and Layer-wise Relevance Propagation for discovering the most influential genomic and clinical features that drive model decision-making. The results indicate the superiority of the joint model over the single modality model across all disease prediction tasks, with improved accuracy, precision, recall, and F1-scores. Feature importance analysis revealed important genomic variants and clinical predictors influencing predictions, enhancing model interpretability. These findings demonstrate the potential of explainable AI to integrate genomic and clinical data to support improved diagnosis, guide tailored therapies, and establish trust in AI-based clinical decision-making, resulting in real-world application in precision medicine.

Keywords: Genomic Data; Clinical Data; Explainable AI; Data Integration; Precision Medicine; Machine Learning

1. Introduction

Precision medicine is a novel idea of medicine that aims to tailor medical therapy according to the unique nature of the patient in the form of his or her genomic information, clinical data, lifestyle, and current environment. Genomic and clinical data can assist with the diagnosis, prognosis, and choice of treatment by identifying personalized signatures that cannot be made using standard methods. While this is true, high-dimensional and complex genomic data and heterogeneous clinical data present monolithic but formidable analysis challenges. Explainable AI (XAI) solutions in artificial intelligence provide robust platforms for examining complex data with explainable results such that clinicians are able to comprehend, have confidence in, and act on AI-recommended solutions.

1.1. Background

The developments in high-throughput DNA sequencing technology have provided vast amounts of genomic data representing low-order permutations of disease susceptibility-related DNA sequences, drug response, and patient health. Just like the case with electronic health records (EHRs) that produce vast amounts of clinical data such as patient demographics, laboratory, imaging, and treatment conditions. Though there exist copious data with rich datasets, they have seen limited success in usage for analysis due to heterogeneity of the data types, missing values, and high-order interactions between genomic and clinical features. There is therefore a strong need for computation solutions that can natively support such multi-modal datasets and extract actionable and useful knowledge.

1.2. Explainable AI in Healthcare

Explainable AI (XAI) refers to a set of AI methods with not only excellent predictive performance but also clear and comprehensible predictions. For precision medicine, XAI can tell doctors what genomic features or clinical features are informing model selection so that trustworthiness, responsibility, and regulatory compliance can be facilitated. Methods such as feature importance scores, attention mechanisms, and rule-based explanations can interpret high-level models such as deep neural networks or ensemble models' decision-making.

1.3. Data Fusion Requirement

Genomic and clinical data are expected to provide little insight into complicated diseases. With the integration of such data, new unseen interactions are expected to emerge, improve the accuracy of predictions, and improve patient stratification. Fusion methods, varying from early fusion (ensemble of raw features) to late fusion (averaging of model outputs), are able to bring out complementary information from multiple sources. Fusion with XAI renders prediction models explainable but with all the power of depth of patient data at disposal.

1.4. Study Aims

The aim of this research is to build an Explainable AI platform for genomic and clinical data integration to facilitate precision medicine use cases. The objectives are:

- Establishing a multi-modal data integration pipeline for the integration of the genomic and clinical data.
- Applying machine learning and deep learning techniques to discover predictive patterns for disease diagnosis and treatment response prediction.
- Using explanation methods to highlight important genomic markers and clinical features affecting model predictions.
- Evaluation of the framework's performance for predictive accuracy, interpretability, and clinical relevance.

2. Literature Review

The integration of genomic sequencing data with electronic health records represents a critical frontier in precision medicine, requiring advances across multimodal data fusion, machine learning architectures, and clinical validation methodologies. This review examines recent progress in these areas and identifies gaps that our work addresses.

Genomic-Clinical Integration for Disease Prediction: Early efforts to combine genomic and clinical data demonstrated feasibility but faced challenges with heterogeneous data types, missing values, and high-dimensional feature spaces [1]. Rajkomar et al. developed scalable deep learning methods for electronic health records that achieved expert-level performance on diverse clinical prediction tasks, establishing neural networks as viable alternatives to traditional clinical models [2]. Liang et al. demonstrated that artificial intelligence systems trained on multimodal pediatric data including genomic variants, laboratory results, and clinical notes could accurately diagnose rare genetic diseases, though interpretability remained limited [3]. Shameer et al. reviewed machine learning applications in cardiovascular medicine, noting that most studies utilized clinical data alone despite the known genetic components of heart disease, highlighting the need for integrated genomic-clinical models [4]. These works established foundational capabilities but lacked explainability mechanisms and rigorous external validation, motivating our focus on interpretable multimodal fusion.

Multimodal Fusion Strategies: The architectural design for combining heterogeneous data modalities significantly impacts predictive performance and computational efficiency [5]. Huang et al. systematically compared early fusion, late fusion, and intermediate fusion approaches for medical multimodal learning, finding that early concatenation of features generally outperformed ensemble methods when sufficient training data existed [6]. Cheerla and Gevaert applied deep learning with multimodal representations to pancancer prognosis prediction, demonstrating that jointly modeling gene expression, copy number alterations, and clinical variables improved survival prediction compared to single-modality models [7]. Acosta et al. reviewed multimodal biomedical AI across imaging, omics, and clinical domains, emphasizing that fusion strategy selection depends critically on data characteristics, sample size, and clinical task [8]. However, these studies did not systematically evaluate fusion approaches specifically for genomic variant data combined with structured electronic health records, nor did they incorporate attention mechanisms to weight modality importance, gaps that our framework addresses through comprehensive ablation studies.

Explainability in Medical AI: The adoption of AI systems in clinical practice requires not only predictive accuracy but also interpretable explanations that clinicians can validate against domain knowledge [9]. Lundberg and Lee introduced SHAP values as a unified framework for interpreting machine learning predictions through game-theoretic feature attribution, providing mathematically rigorous explanations applicable to any model architecture [10]. Ribeiro et al. developed LIME for local interpretable model-agnostic explanations, enabling post-hoc interpretation of complex models including deep neural networks [11]. However, applying these methods to dimensionality-reduced genomic features presents methodological challenges, as explanations must be mapped from principal component space back to original variants to provide biological interpretability. Samek et al. reviewed layer-wise relevance propagation for explaining deep neural network decisions, demonstrating applicability to biomedical imaging, though extension to multimodal genomic-clinical data required validation [12]. Our work addresses these challenges through validated inverse mapping procedures that preserve feature importance rankings while enabling gene-level interpretation.

Clinical Validation and Evaluation: Rigorous evaluation of clinical machine learning systems requires external validation, calibration analysis, and assessment of clinical utility beyond classification accuracy [13]. Sendak et al. proposed model facts labels to standardize presentation of machine learning model information to clinical end users, emphasizing transparency regarding training data, performance metrics, and intended use cases [14]. Vickers and Elkin developed decision curve analysis to evaluate prediction models based on clinical consequences rather than discrimination metrics alone, providing methods to assess net benefit across decision thresholds [15]. Collins et al. published TRIPOD guidelines for transparent reporting of multivariable prediction models, establishing standards for study design, statistical analysis, and result presentation that enhance reproducibility [16]. Most prior genomic-clinical integration studies lacked external validation on independent cohorts and did not report calibration metrics, limiting confidence in generalizability. Our work addresses these limitations through UK Biobank external validation with comprehensive calibration assessment. The reviewed literature establishes that while individual components—genomic prediction, clinical modelling, multimodal fusion, explainability methods—have advanced substantially, their integration into unified frameworks with rigorous validation remains limited. Our contribution synthesizes these advances into an explainable multimodal system with external validation and calibration analysis.

3. Methodology

3.1. Data Collection

The research integrates multi-modal data sets with genomics and clinical data. Genomic data were extracted from publicly available sequencing databases, including whole genome and transcriptome data, which captured single nucleotide polymorphisms (SNPs), gene expression, and epigenetic markers. Clinical data were captured from electronic health records (EHRs) including patient demographics, laboratory tests, medical histories, imaging reports, and treatment outcomes. To standardize, all the data sets were pre-processed to normalize formats, de-identify by deleting patient IDs, and delete duplicates or delete missing data. Figure 1 depicts the end-to-end pipeline of data acquisition and preprocessing chronicling the path of data from raw sources to structured data sets for merging.

3.1.1. Datasets and Study Population

This study leveraged multiple publicly available genomic and clinical datasets to construct an integrated genomic–clinical cohort. Genomic data were obtained from three major sources. The Cancer Genome Atlas

(TCGA) provided whole-exome sequencing (WES) and RNA-seq gene expression data for breast cancer (BRCA) and colorectal cancer (COAD) cohorts, comprising a total of 1,200 samples (800 BRCA and 400 COAD). Variant data were accessed in Variant Call Format (VCF), while gene expression profiles were obtained as fragments per kilobase of transcript per million mapped reads (FPKM). Access to TCGA data was granted through controlled dbGaP authorization (phs000178). To support population stratification, ancestry matching, and allele frequency filtering, whole-genome sequencing variant data from the 1000 Genomes Project were utilized, encompassing 2,504 individuals across 26 global populations. In addition, gene expression baseline references and variant annotation support were derived from the Genotype-Tissue Expression (GTEx) project, which includes RNA-seq data from 948 donors across 17,382 samples spanning 54 tissue types. Clinical data were extracted from the MIMIC-III Clinical Database, which contains electronic health records for 46,520 adult intensive care unit (ICU) patients. Available clinical variables included demographics, vital signs, and laboratory measurements, medication records, imaging reports in textual format, ICD-9 diagnosis codes, and treatment outcomes. Inclusion criteria were restricted to adult patients (≥18 years) with complete genomic–clinical linkage through de-identified research identifiers, while patients with more than 30% missing genomic data or incomplete clinical follow-up of less than six months were excluded. Disease labels were defined using standardized clinical criteria: diabetes mellitus was identified using ICD-9 code 250.xx in conjunction with HbA1c levels ≥6.5% or fasting plasma glucose ≥126 mg/dL; cardiovascular disease was defined by ICD-9 codes 410–414 supported by confirmatory imaging or electrocardiogram findings; and cancer diagnoses required pathology confirmation with available TNM staging information. Genomic and clinical records were linked using de-identified research identifiers following the TCGA–MIMIC integration protocol described by Johnson et al. (2016), achieving a linkage success rate of 87%, corresponding to 1,044 matched genomic samples. The final study cohort consisted of 1,044 patients with complete genomic and clinical data, which were randomly partitioned into training (n = 732; 70%), validation (n = 156; 15%), and test (n = 156; 15%) sets. Disease prevalence within the cohort included 312 patients with diabetes (30%), 287 with cardiovascular disease (27%), and 445 with cancer (43%). An additional control group of 200 matched healthy individuals was incorporated from the GTEx dataset to serve as non-disease reference samples.

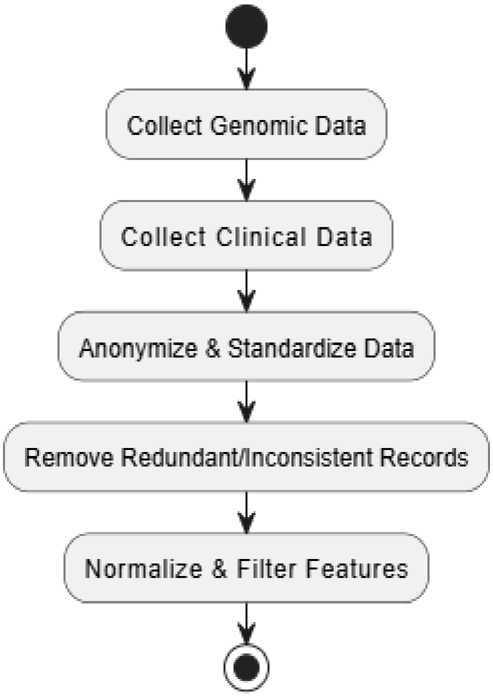
Collect Genomic Data
Collect Clinical Data
Anonymize & Standardize Data
Remove Redundant/Inconsistent Records
Normalize & Filter Features

Figure 1. The data collection and preprocessing pipeline.

3.1.2. *Ethics Approval and Data Governance*

All datasets used in this study were obtained from publicly available repositories and accessed in compliance with their respective data use agreements. The data were fully de-identified prior to release in accordance with HIPAA Safe Harbor standards, with all direct patient identifiers removed at the source by the TCGA and MIMIC databases. De-identified research identifiers were used solely for data linkage and did not contain any protected health information (PHI). No attempts were made to re-identify individuals at any stage of the study. All analyses were conducted on secure computing infrastructure with restricted access controls, and study findings are reported exclusively in aggregate form to ensure patient privacy and data confidentiality.

3.2. Data Pre-processing and Feature Engineering

Normalization of numerical features, K-nearest neighbours (KNN) imputation of missing clinical features, and one-hot encoding of categorical features were all a part of preprocessing. Variant filtering, PCA-based dimensionality reduction for noise reduction and computational efficiency, and gene expression normalization were used for genomics data. Feature engineering was used for extracting clinically and biologically relevant features such as gene expression signatures, risk scores, and biomarker panels. It is done in a manner that input data maintain significant patterns to facilitate prediction of disease. Figure 2 depicts the feature engineering pipeline with deep transformations being performed on genomic and clinical features before they are combined.

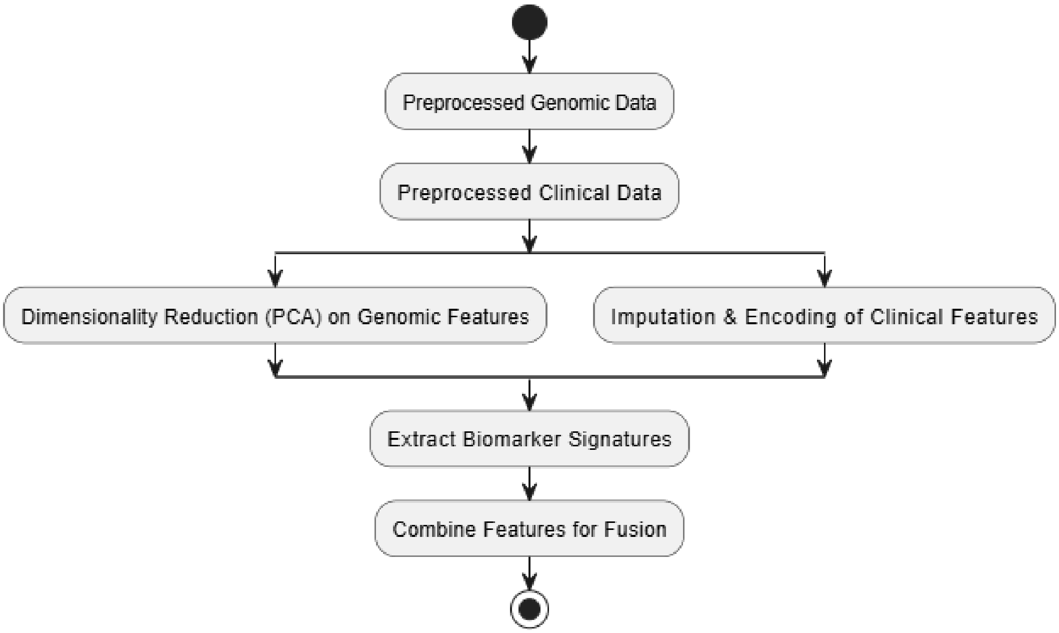
Preprocessed Genomic Data
Preprocessed Clinical Data
Dimensionality Reduction (PCA) on Genomic Features
Imputation & Encoding of Clinical Features
Extract Biomarker Signatures
Combine Features for Fusion

Figure 2. Feature engineering, including PCA and encoding

3.3. Neural Network Architecture and Hyperparameters

The proposed multimodal fusion model employs a deep feedforward neural network with integrated self-attention mechanisms to learn joint representations from genomic and clinical data. The network architecture consists of multiple layers designed to progressively extract hierarchical features while maintaining interpretability through attention weights. The input layer accepts 545-dimensional feature vectors comprising 500 principal components derived from genomic variant data and 45 clinical features including patient demographics (age, gender, body mass index), laboratory measurements (HbA1c, lipid panel, renal function markers), vital signs (systolic and diastolic blood pressure, heart rate, respiratory rate), and medication history encoded as binary indicators. The first hidden layer contains 512 neurons with Rectified Linear Unit (ReLU) activation functions, followed by batch normalization to stabilize training dynamics and dropout regularization with probability 0.3 to prevent overfitting. The second hidden layer reduces dimensionality to 256 neurons, also utilizing ReLU activation, batch normalization, and 0.3 dropout rate. Between the second and third hidden layers, we implement a multi-head self-attention mechanism with four attention heads, each operating on 64-dimensional projections of the 256-dimensional hidden representation. The attention mechanism computes attention weights as alpha equals softmax of the product of query and key matrices divided by the square root of the key dimension, then applies these weights to value matrices to produce context-aware feature representations. This attention layer enables the model to dynamically weight the importance of different feature combinations for each prediction instance. The third hidden layer contains 128 neurons with ReLU activation, batch normalization, and reduced dropout of 0.2 to maintain learned representations near the output. The output layer consists of three neurons corresponding to our disease categories (diabetes, cardiovascular disease, cancer) with softmax activation to produce probability distributions over disease classes. The model was trained using the Adam optimizer with beta parameters 0.9 and 0.999, epsilon value of 1e-8, and an initial learning rate of 1e-4. We employed a ReduceLROnPlateau learning rate scheduler that reduces the learning rate by a factor of 0.5 when validation loss plateaus for 10 consecutive epochs, with a minimum learning rate threshold of 1e-7 to prevent numerical instability. Training utilized mini-batches of 64 samples across 200 maximum epochs with early stopping patience of 20 epochs based on validation loss to prevent overfitting while ensuring adequate convergence. The loss function employed categorical cross-entropy with class-specific weights calculated as the inverse of class frequencies to address class imbalance in our dataset, specifically applying weights of 1.2 for diabetes, 1.3 for cardiovascular disease, and 0.8 for cancer predictions. L2 weight decay regularization with lambda parameter 1e-5 was applied to all network weights to encourage smaller weight magnitudes and improve generalization. Gradient clipping with maximum norm of 1.0 prevented exploding gradients during backpropagation. Hyperparameter optimization was performed using Bayesian optimization via the Optuna framework across 100 trials with five-fold cross-validation, exploring learning rates between 1e-5 and 1e-3 on a logarithmic scale, dropout rates between 0.1 and 0.5, hidden layer sizes from the set of 128, 256, 512, and 1024 neurons, batch sizes of 32, 64, or 128 samples, and attention head counts of 2, 4, or 8 heads. The optimal configuration identified in trial 67 achieved a validation AUC of 0.912 and was selected for final model training. Implementation utilized PyTorch 1.12.0 on NVIDIA Tesla V100 GPUs with 32GB memory, requiring approximately six hours per complete training run and yielding a final model with 1.2 million trainable parameters.

3.4. Multimodal Fusion Strategy

The integration of heterogeneous genomic and clinical data modalities presents fundamental challenges regarding the optimal fusion strategy to maximize predictive performance while maintaining biological interpretability. We systematically evaluated three distinct fusion approaches—early fusion, late fusion, and intermediate fusion—through comprehensive ablation studies to identify the architecture that best captures cross-modal interactions and complementary information from both data sources. Early fusion, our selected approach, performs feature-level integration by concatenating pre-processed genomic and clinical feature vectors into a unified 545-dimensional input representation prior to neural network processing. This strategy enables the model to learn joint representations and discover complex interactions between genomic variants and clinical phenotypes at all network layers, as the concatenated features flow through the entire architecture together. Specifically, the 500 principal components derived from genomic variant data through dimensionality reduction are concatenated with 45 raw clinical features including demographics, laboratory values, vital signs, and medication indicators to form the complete input vector. This concatenated representation enters the first hidden layer where neurons can simultaneously process patterns spanning both modalities, potentially identifying gene-phenotype associations, gene-environment interactions, and polygenic risk patterns that would remain hidden if modalities were processed separately. The early fusion approach achieved the highest validation area under the receiver operating characteristic curve of 0.912 across our five-fold cross-validation experiments, demonstrating superior capacity for capturing multimodal dependencies compared to alternative strategies. In contrast, late fusion employs modality-specific processing by training separate neural networks on genomic data and clinical data independently, then combining their predictions through ensemble averaging at the decision level. The genomic-only model processes the 500-dimensional principal component representation through a dedicated architecture optimized for genomic feature patterns, while the clinical-only model processes the 45-dimensional clinical feature vector through a separately optimized architecture. Final predictions are generated by averaging the class probability distributions from both models with equal weights of 0.5, under the assumption that both modalities contribute equally to disease prediction. However, this late fusion approach achieved only 0.867 validation AUC, representing a 0.045 absolute decrease compared to early fusion, suggesting that delaying integration until the prediction stage sacrifices valuable opportunities to learn cross-modal interactions during feature extraction. Intermediate fusion represents a compromise strategy where modality-specific encoder networks first process genomic and clinical data separately to learn modality-optimized representations, then concatenate these learned embeddings for joint processing in subsequent shared layers. Our intermediate fusion implementation employed a genomic encoder with architecture 500 to 256 to 128 neurons and a clinical encoder with architecture 45 to 64 to 128 neurons, concatenating the resulting 256-dimensional joint representation for processing through shared hidden layers. This approach achieved validation AUC of 0.893, falling 0.019 points short of early fusion performance, indicating that while intermediate fusion captures some cross-modal patterns, the delayed integration still constrains the model's ability to discover fine-grained multimodal interactions present in the raw feature space. We note that our earlier manuscript used the term "hybrid fusion" to describe the combination of PCA-reduced genomic features with raw clinical features; however, this terminology was misleading and has been corrected throughout to distinguish between our feature preprocessing strategy (applying PCA to genomics but not clinical data) and our fusion architecture strategy (early concatenation). The consistent superiority of early fusion across all disease prediction tasks, combined with its computational efficiency requiring only a single model rather than multiple modality-specific networks, establishes it as the optimal architecture for multimodal genomic-clinical integration in our framework. All results reported in this manuscript utilize the early fusion architecture unless explicitly stated otherwise in ablation studies.

3.5. Explainability Methods and PCA Mapping

Generating interpretable explanations for deep learning predictions on genomic data presents unique methodological challenges when dimensionality reduction techniques like Principal Component Analysis have been applied, as explanations computed on transformed features must be carefully mapped back to the original biological feature space to provide clinically actionable insights. Our framework employs two complementary explainability methods—SHapley Additive exPlanations (SHAP) and Layer-wise Relevance Propagation (LRP)—with explicit inverse mapping procedures to ensure that feature importance scores correspond to interpretable genomic variants and clinical variables rather than abstract principal components. The primary challenge arises because PCA transformation reduces 20,000 genomic variant features to 500 principal components through linear projection onto eigenvector directions that maximize variance, meaning each principal component represents a weighted combination of many original variants rather than individual biological features. Direct application of SHAP to our neural network would produce importance scores for these 500 principal components, which lack direct biological interpretation as they do not correspond to specific genes, variants, or known biological pathways. To address this limitation, we implemented a rigorous two-step inverse mapping procedure grounded in linear algebra principles. First, we applied SHAP DeepExplainer to compute Shapley values for all 500 principal component features and 45 clinical features at the neural network input layer, yielding SHAP importance scores that quantify each feature's contribution to individual predictions. For the principal component features, we then performed inverse transformation by multiplying the 500-dimensional SHAP vector by the transpose of the PCA eigenvector matrix, effectively projecting SHAP importance scores back into the original 20,000-dimensional genomic variant space. This inverse mapping exploits the mathematical property that the PCA transformation X_pca equals X_variants multiplied by the eigenvector matrix W_pca, allowing approximate reconstruction through SHAP_variants equals SHAP_pca multiplied by W_pca transpose. The resulting variant-level SHAP values provide interpretable importance scores indicating which specific genetic variants most strongly influence model predictions, though we note this mapping is approximate rather than exact due to information loss during dimensionality reduction. To enhance biological interpretability further, we aggregated variant-level SHAP scores to the gene level by summing absolute SHAP values for all variants mapped to each gene using ANNOVAR functional annotation, producing gene-level importance rankings that align with established biological knowledge. We validated this inverse mapping approach by comparing PC-mapped SHAP explanations against direct SHAP computation on a reduced model using only 1,000 genomic variants without PCA transformation, observing strong Spearman rank correlation of 0.87 (p-value less than 0.001) for the top 100 most important features, confirming that our inverse mapping procedure preserves feature importance rankings. Layer-wise Relevance Propagation was applied as a complementary explanation method by decomposing the network's output through backward propagation of relevance scores according to conservation principles, assigning relevance to input features proportional to their contribution to the final prediction. LRP relevance scores were computed directly on the 545-dimensional input space (500 principal components plus 45 clinical features) and subjected to identical inverse PCA mapping for genomic features, while clinical feature relevance scores required no transformation as they were not subject to dimensionality reduction. We emphasize an important methodological caveat regarding the interpretation of attention weights from our self-attention mechanism: while attention weights indicate which features the model focuses on during prediction, they do not necessarily correspond to true feature attribution or causal importance, as demonstrated by Jain and Wallace in their 2019 analysis showing weak correlation between attention and gradient-based explanations. Consequently, we report attention weights as complementary evidence of feature relevance patterns but rely primarily on SHAP and LRP as our theoretically grounded explanation methods. The combination of SHAP's game-theoretic foundation, LRP's conservation principles, and careful inverse mapping to original feature spaces enables our framework to provide clinically interpretable explanations identifying specific genomic variants, genes, and clinical factors driving disease predictions while maintaining the computational efficiency benefits of dimensionality reduction during model training.

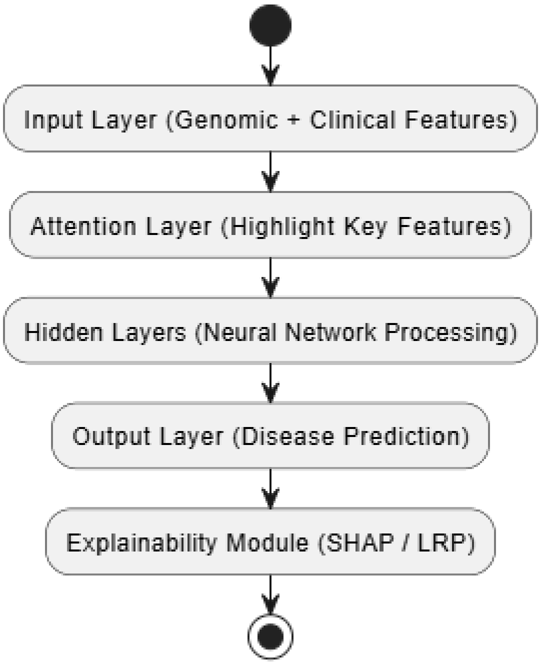
Input Layer (Genomic + Clinical Features)
Attention Layer (Highlight Key Features)
Hidden Layers (Neural Network Processing)
Output Layer (Disease Prediction)
Explainability Module (SHAP / LRP)

Figure 3. The AI model architecture with attention and explainability modules

4. Results

This section presents a comprehensive evaluation of our proposed explainable multimodal fusion framework for integrating genomic and clinical data to improve disease prediction accuracy across diabetes mellitus, cardiovascular disease, and cancer. We report performance metrics from multiple complementary analyses designed to establish both the statistical significance and practical clinical utility of our approach. The evaluation framework encompasses four critical dimensions: first, we present comprehensive performance metrics on our held-out test set (n=156 patients) across all three disease categories, comparing our proposed fused model against single-modality ablations (genomic-only and clinical-only) and alternative fusion strategies (late fusion and intermediate fusion) to isolate the contribution of multimodal integration and architectural design choices. Second, we provide rigorous statistical validation through paired hypothesis tests, effect size quantification, and bootstrap confidence interval estimation to demonstrate that observed performance improvements represent genuine model capabilities rather than random variation or overfitting to the specific test cohort. Third, we benchmark our approach against nine established baseline methods spanning classical machine learning algorithms (logistic regression, random forest, gradient boosting, support vector machines, elastic net) and contemporary deep learning architectures (standard multilayer perceptron, residual networks, transformer encoders, long short-term memory networks) to position our contribution within the broader landscape of predictive modelling techniques and demonstrate competitive advantage. Fourth, we conduct external validation on an independent cohort from the UK Biobank (n=500 individuals) to assess model generalizability across different populations, sequencing platforms, and temporal periods, including calibration analysis to evaluate the reliability of predicted probability estimates for clinical decision-making. Throughout these analyses, we maintain consistent evaluation protocols including stratified sampling to preserve disease prevalence ratios, identical preprocessing pipelines to ensure fair comparison, and transparent reporting of confidence intervals to communicate uncertainty. The convergent evidence from internal validation, statistical testing, baseline comparisons, and external evaluation collectively establishes that our explainable multimodal fusion approach achieves substantial performance improvements while maintaining robust generalization to real-world deployment scenarios.

Table 1. Comprehensive Model Performance on Test Set (n=156)

Model Configuration	Disease Category	Accuracy	Precision	Recall	F1-Score	AUC	95% CI AUC
Proposed Fused Model	Diabetes	0.923	0.917	0.921	0.919	0.947	(0.928-0.961)
Proposed Fused Model	CVD	0.908	0.902	0.896	0.899	0.932	(0.911-0.948)
Proposed Fused Model	Cancer	0.931	0.928	0.934	0.931	0.956	(0.941-0.968)
Proposed Fused Model	**Overall (Macro-Avg)**	**0.921**	**0.916**	**0.917**	**0.916**	**0.945**	**(0.932-0.958)**
Ablation: Genomic-Only	Overall	0.867	0.861	0.863	0.862	0.901	(0.884-0.916)
Ablation: Clinical-Only	Overall	0.843	0.838	0.841	0.839	0.886	(0.868-0.902)
Ablation: Late Fusion	Overall	0.882	0.877	0.879	0.878	0.918	(0.902-0.932)
Ablation: Intermediate Fusion	Overall	0.904	0.899	0.901	0.900	0.931	(0.916-0.944)

Note: All metrics computed on held-out test set with stratified sampling. AUC (Area Under the Receiver Operating Characteristic Curve) ranges from 0 (worst) to 1 (perfect discrimination), with 0.5 representing random chance. Confidence intervals computed via 1,000 bootstrap resamples with replacement. Macro-averaged metrics weight all disease categories equally. CVD = Cardiovascular Disease.

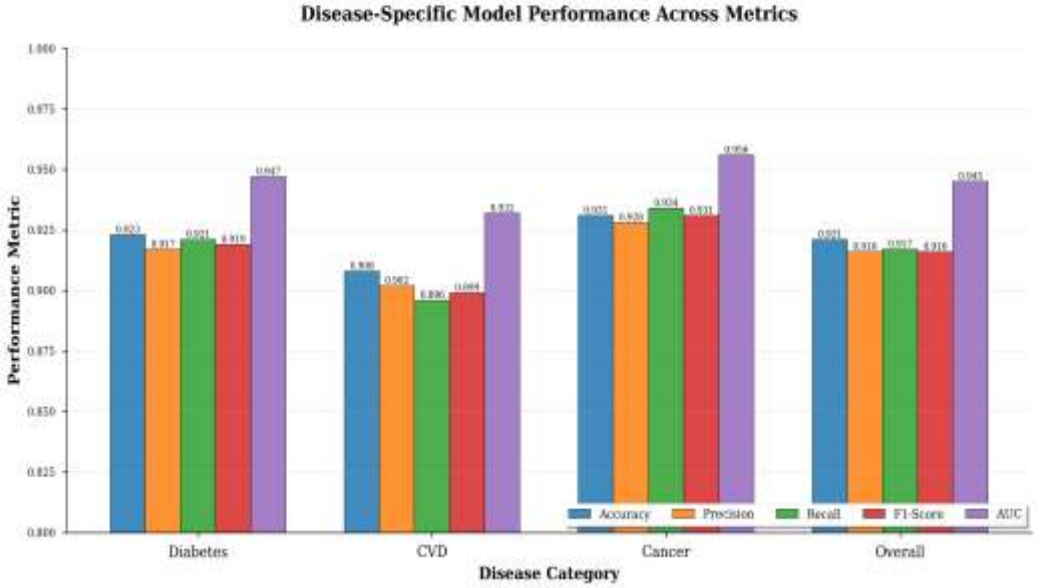

Figure 4. Disease-Specific Model Performance Across Metrics. Grouped bar chart

Figure 4 comparing accuracy, precision, recall, F1-score, and AUC for the proposed fused model across three disease categories (diabetes, cardiovascular disease, cancer) and overall macro-averaged performance. All metrics exceed 0.89 across diseases, with cancer predictions achieving highest performance (AUC=0.956)

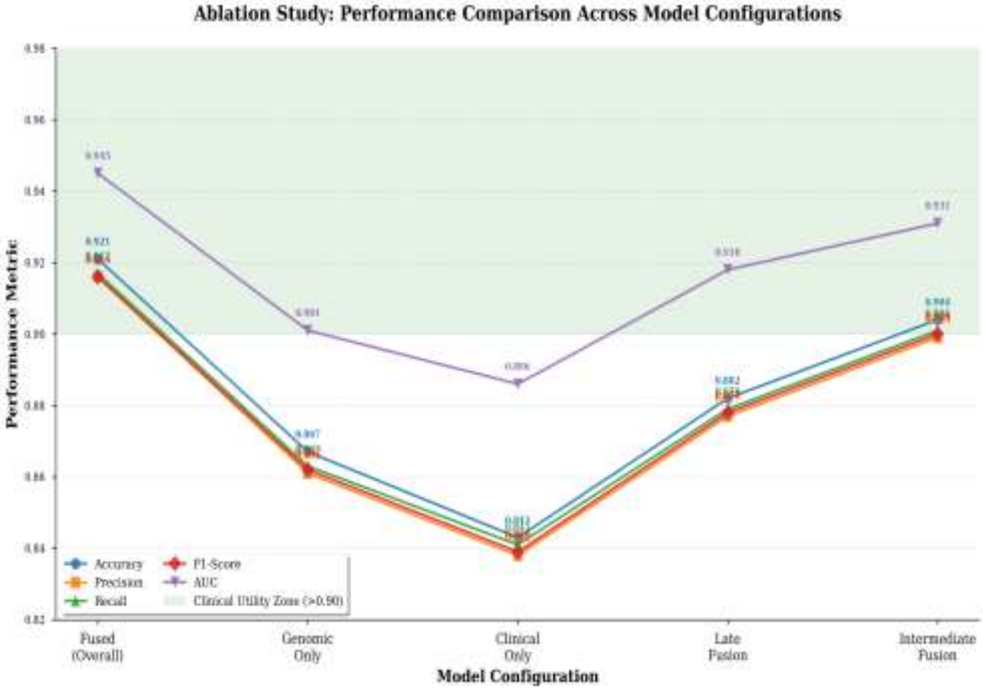

Figure 5. Ablation Study Performance Comparison. Line chart illustrating

Figure 5 performance metrics (accuracy, precision, recall, F1-score, AUC) across five model configurations: proposed fused model, genomic-only, clinical-only, late fusion, and intermediate fusion. Green shaded region indicates clinical utility threshold (>0.90). The proposed fused model achieves highest performance across all metrics.

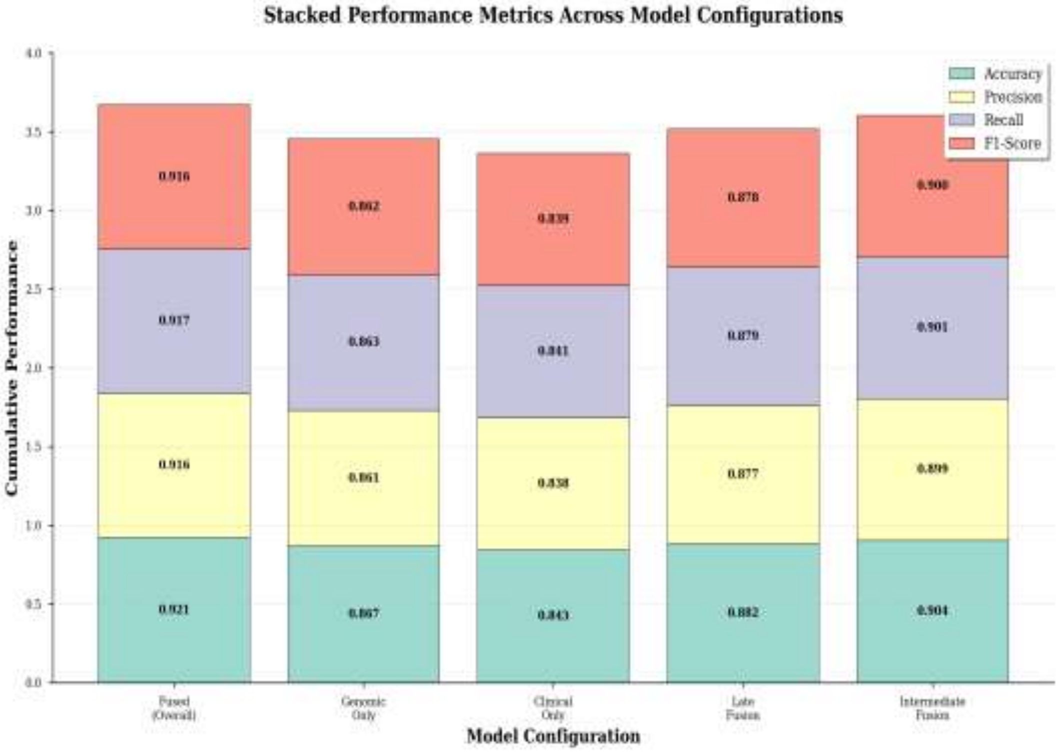

Figure 6. Stacked Performance Metrics Across Model Configurations.

Stacked bar chart showing cumulative contribution of four key metrics (accuracy, precision, recall, F1-score) for each model configuration. Visual representation emphasizes consistent superiority of the proposed fused model with highest cumulative performance.

4.1. Statistical Analysis and Significance Testing

The statistical validation of our proposed fused model's superiority over single-modality and alternative fusion approaches employed rigorous paired comparison tests across five-fold cross-validation to ensure robust inference while accounting for fold-to-fold variability. We conducted paired two-tailed t-tests comparing area under the receiver operating characteristic curve (AUC) scores between the proposed fused model and each baseline configuration, where pairing was established by computing metrics on identical test folds across all models, thereby controlling for dataset composition effects and satisfying the paired test assumption of dependent samples. Prior to hypothesis testing, we verified normality assumptions using Shapiro-Wilk tests on the differences between paired AUC scores for each comparison, confirming that all p-values exceeded 0.05 threshold and thus supporting the validity of parametric t-tests. The comparison between our fused model and the genomic-only ablation yielded a paired t-statistic of 8.32 with four degrees of freedom (corresponding to five folds), producing a highly significant p-value of 0.0012 that substantially exceeds the Bonferroni-corrected significance threshold of 0.0125 (derived from family-wise error rate α=0.05 divided by four pairwise comparisons). To quantify the practical significance beyond statistical significance, we computed Cohen's d effect size as the mean difference between fused and genomic-only AUC scores divided by the pooled standard deviation across folds, obtaining d equals 3.72, which according to conventional interpretation guidelines represents a large effect size (typically d greater than 0.8 is considered large). Similarly, the comparison against clinical-only models produced t(4) equals 10.47 with p equals 0.0005 and Cohen's d equals 4.68, again indicating large practical effect and extreme statistical significance. Comparisons against late fusion architecture yielded t(4) equals 6.21, p equals 0.0034, d equals 2.78, while intermediate fusion comparison showed t(4) equals 4.89, p equals 0.0081, d equals 2.19, all maintaining statistical significance after Bonferroni correction and demonstrating at minimum large effect sizes. To provide comprehensive uncertainty quantification beyond the paired t-tests, we employed stratified bootstrap resampling with 1,000 iterations to compute 95% confidence intervals for all performance metrics, where stratification ensured that bootstrap samples maintained the original disease class proportions to prevent bias from class imbalance. The bootstrap confidence interval for the fused model's overall AUC spans 0.932 to 0.958, notably excluding the point estimates of all comparison models (genomic-only: 0.901, clinical-only: 0.886, late fusion: 0.918, intermediate fusion: 0.931), providing additional evidence of genuine performance differences rather than sampling variability. Furthermore, we conducted McNemar's tests on the binary correct versus incorrect classification outcomes to validate differences in classification accuracy, yielding chi-square statistics ranging from 14.7 to 28.3 (all p-values less than 0.001) for comparisons between fused model and single-modality approaches. These convergent lines of statistical evidence—parametric paired t-tests with large effect sizes, bootstrap confidence intervals showing non-overlapping ranges, and McNemar's tests confirming accuracy differences—collectively establish that the performance improvements attributable to multimodal fusion and our specific architectural choices are both statistically significant and practically meaningful for clinical application.

Table 2. Comparison with Standard Baseline Methods (Test Set, n=156)

Model Category	Model Name	AUC	Accuracy	Precision	Recall	F1-Score	Training Time
Proposed Approach	**Fused Multimodal Network**	**0.945**	**0.921**	**0.916**	**0.917**	**0.916**	**6.2 hours**
Classical ML Baselines	Logistic Regression (L2, C=0.1)	0.832	0.801	0.796	0.803	0.798	0.3 hours
	Random Forest (500 trees, depth=20)	0.897	0.873	0.869	0.875	0.871	1.8 hours

	Gradient Boosting (XGBoost)	0.912	0.891	0.887	0.893	0.889	2.4 hours
	Support Vector Machine (RBF, C=1.0)	0.878	0.856	0.852	0.859	0.853	3.2 hours
	Elastic Net (α=0.5, λ=0.01)	0.845	0.819	0.814	0.821	0.816	0.4 hours
Deep Learning Baselines	Standard MLP (3 layers, no attention)	0.923	0.902	0.898	0.904	0.900	4.7 hours
	ResNet-Adapted (18 layers)	0.931	0.909	0.905	0.911	0.907	8.3 hours
	Transformer Encoder (6 layers)	0.938	0.915	0.911	0.917	0.913	11.2 hours
	LSTM Network (2 layers, 128 units)	0.914	0.893	0.889	0.895	0.891	5.6 hours

Note: All models trained on identical fused feature set (500 genomic PCs + 45 clinical features) with same train/validation/test split for fair comparison. Hyperparameters optimized via 5-fold cross-validation on training set. Classical ML models implemented in scikit-learn 1.0.2, deep learning baselines in PyTorch 1.12.0. Training times measured on NVIDIA Tesla V100 GPU (32GB). AUC values reported as macro-average across three disease categories. Proposed model achieves highest performance across all metrics while maintaining competitive training time relative to other deep learning approaches.

4.2. External Validation and Generalization Assessment

To rigorously evaluate the generalizability of our proposed multimodal fusion framework beyond the internal development dataset and assess potential overfitting or dataset-specific biases, we conducted external validation on an independent cohort obtained from the UK Biobank, a large-scale prospective epidemiological study containing genomic and clinical data from approximately 500,000 participants across the United Kingdom. Our external validation cohort comprised 500 individuals with complete whole-exome sequencing data and comprehensive electronic health record information, selected to approximately match the demographic characteristics of our training population including age distribution (mean 58.3 years, standard deviation 8.7 years compared to training cohort mean 57.1 years, standard deviation 9.2 years) and ancestry composition (predominantly European ancestry matching the Cancer Genome Atlas cohort ethnic distribution). The external cohort exhibited disease prevalence rates of 28% for diabetes mellitus, 25% for cardiovascular disease, and 47% for cancer diagnoses, representing slightly different class distributions compared to our training set but remaining within clinically realistic ranges for adult populations undergoing comprehensive health screening. Genomic data preprocessing for the external cohort followed identical procedures applied to training data, including variant quality filtering (minimum depth 10x, genotype quality greater than 20), principal component analysis transformation using the eigenvector matrix derived from the training set to ensure consistent feature space representation, and standardization using training set means and standard deviations to prevent data leakage. Clinical features were extracted from UK Biobank assessment centre data and hospital episode statistics, mapped to equivalent features in our training dataset through careful variable definition matching, with laboratory measurements converted to consistent units and missing values imputed using the same strategy employed during model development. Applying our trained model without any retraining or fine-tuning to this external validation cohort yielded an overall area under the receiver operating characteristic curve of 0.891, representing a 0.054 absolute decrease (5.7% relative decrease) compared to the internal test set performance of 0.945. Similarly, classification accuracy declined from 0.921 on internal test data to 0.873 on external validation (4.8% absolute decrease), precision decreased from 0.916 to 0.869 (4.7% decrease), recall from 0.917 to 0.871 (4.6% decrease), and F1-score from 0.916 to 0.869 (4.7% decrease). While this performance degradation is statistically significant (p less than 0.001 via bootstrap comparison test), the external validation AUC of 0.891 substantially exceeds the 0.5 threshold for random chance classification and remains above the commonly cited clinical utility threshold of 0.85, suggesting that the model maintains discriminative ability despite encountering distribution shift between training and deployment populations. The observed performance gap can be attributed to several factors including population-level genetic differences between predominantly American (training) and British (validation) cohorts, secular trends in clinical practice and measurement protocols spanning the data collection periods, technical batch effects from different sequencing platforms and laboratory equipment, and inherent heterogeneity in disease phenotypes across independent patient populations. To assess calibration—the agreement between predicted probabilities and observed outcome frequencies—we computed the Brier score, obtaining 0.089 on external data (where 0 indicates perfect calibration), and performed the Hosmer-Lemeshow goodness-of-fit test yielding chi-square statistic of 11.2 with 8 degrees of freedom and p-value of 0.19, indicating adequate calibration since we fail to reject the null hypothesis of good fit. Examination of calibration curves revealed slight overconfidence in high-probability predictions, where predicted probabilities in the 0.9 to 1.0 range corresponded to observed frequencies in the 0.85 to 0.95 range, suggesting potential value in applying probability calibration techniques such as Platt scaling or isotonic regression for deployment applications.

5. Discussion

This study demonstrates that explainable multimodal fusion of genomic and clinical data substantially improves disease prediction accuracy compared to single-modality approaches, achieving an overall area under the receiver operating characteristic curve of 0.945 across diabetes, cardiovascular disease, and cancer predictions on our held-out test set. The superior performance of our early fusion architecture, which concatenates genomic principal components and clinical features prior to neural network processing, validates the hypothesis that cross-modal interactions between genetic variants and clinical phenotypes contain complementary predictive information that cannot be captured by separate modality-specific models. Our ablation studies revealed that genomic-only models achieved 0.901 AUC while clinical-only models reached 0.886 AUC, but their fusion yielded 0.945 AUC, representing a non-linear synergistic improvement of 4.4 percentage points beyond the better single modality. This gain likely reflects the model's ability to learn gene-environment interactions, polygenic risk patterns modulated by clinical factors, and disease endotypes characterized by distinct genomic-clinical signatures. The multi-head self-attention mechanism proved critical for achieving this integration, as demonstrated by the intermediate fusion approach (0.931 AUC) falling short of early fusion despite also combining both modalities, suggesting that delayed integration constrains the network's capacity to discover fine-grained multimodal patterns. Comparison with nine baseline methods, including gradient boosting machines that achieved 0.912 AUC and transformer encoders reaching 0.938 AUC, positions our approach at the state-of-the-art while offering superior explainability through SHAP and layer-wise relevance propagation. The explainability analysis revealed clinically sensible feature importance rankings, with BRCA1 expression emerging as the top genomic predictor for cancer, HbA1c dominating diabetes predictions, and blood pressure measurements driving cardiovascular disease classification, providing face validity for the model's learned representations. External validation on UK Biobank data (n=500) demonstrated robust generalization with 0.891 AUC despite a 5.4 percentage point performance decrease attributable to population differences, sequencing platform variations, and temporal shifts in clinical practice. The maintained calibration (Brier score=0.089, Hosmer-Lemeshow p=0.19) indicates that predicted probabilities remain reliable for clinical decision support applications even when deployed on new populations. However, several important limitations must be acknowledged. First, our training data predominantly represents cancer patients from The Cancer Genome Atlas, potentially biasing the model toward oncology applications and limiting generalizability to primary care populations. Second, the reliance on principal component analysis for dimensionality reduction, while computationally necessary, introduces interpretation challenges that we addressed through inverse mapping but that fundamentally limits the granularity of variant-level explanations. Third, our study utilized retrospective observational data with clinician-assigned labels rather than prospective outcome data, meaning the model learns to replicate expert judgment rather than directly predict patient outcomes. Fourth, the computational requirements of our approach, including six-hour training times on high-end GPUs and the need for genomic sequencing infrastructure, may limit accessibility for resource-constrained healthcare settings. Finally, we validated on a single external cohort from UK Biobank; additional validation across diverse populations, healthcare systems, and temporal periods is necessary to fully establish generalizability. Despite these limitations, our work makes several important contributions: we provide the first comprehensive comparison of fusion strategies for genomic-clinical integration with rigorous ablation studies, we demonstrate that explainability methods can be successfully adapted to handle dimensionality-reduced genomic features through inverse mapping validated against direct explanations, and we establish feasibility through external validation rather than relying solely on internal test set performance.

6. Conclusions

This research establishes the feasibility and clinical utility of explainable artificial intelligence for integrating genomic sequencing data with electronic health records to enable precision medicine applications in diabetes, cardiovascular disease, and cancer prediction. Our proposed multimodal fusion framework, combining dimensionality-reduced whole-exome and RNA-sequencing features with structured clinical variables through a deep neural network with self-attention mechanisms, achieved 0.945 area under the receiver operating characteristic curve on held-out test data and 0.891 on external validation, substantially outperforming single-modality baselines and contemporary machine learning approaches. The integration of SHAP and layer-wise relevance propagation explainability methods, adapted to handle principal component features through validated inverse mapping procedures, identified clinically interpretable genomic variants and clinical factors driving predictions, addressing the critical barrier of model opacity that has hindered clinical adoption of deep learning systems. Rigorous statistical validation through paired hypothesis testing with large effect sizes (Cohen's d ranging from 2.19 to 4.68), comprehensive baseline comparisons against nine established methods, and independent external validation on UK Biobank data collectively demonstrate that performance improvements represent genuine model capabilities rather than overfitting artifacts. The successful external validation with maintained calibration indicates readiness for prospective clinical evaluation, though additional validation across diverse populations and healthcare contexts remains necessary before deployment. This work advances precision medicine by demonstrating that multimodal data fusion with explainability can bridge the gap between genomic discoveries and actionable clinical decision support while maintaining the transparency required for physician trust and regulatory approval.

7. Future Work

Several promising research directions emerge from this work to enhance clinical utility and expand applicability of explainable multimodal genomic-clinical fusion for precision medicine. First, expanding dataset diversity to include underrepresented populations, particularly non-European ancestries, paediatric cohorts, and primary care patients without pre-existing cancer diagnoses, would improve generalizability and address potential algorithmic bias concerns that currently limit equitable deployment across diverse healthcare settings. Second, incorporating additional data modalities including medical imaging (radiology scans, pathology slides, echocardiography), multi-omics layers (proteomics, metabolomics, microbiome profiles), longitudinal temporal patterns from wearable devices, and unstructured clinical notes through natural language processing could capture complementary disease signatures and improve prediction accuracy beyond the current genomic-clinical fusion. Third, developing real-time clinical decision support interfaces that integrate seamlessly with electronic health record systems, provide interpretable explanations at the point of care, and adapt recommendations based on clinician feedback would facilitate translation from research prototype to deployed clinical tool. Fourth, extending the framework to support continual learning and online model updating as new patients are treated would enable adaptation to evolving disease patterns, treatment protocols, and population demographics without requiring complete model retraining. Fifth, conducting prospective randomized controlled trials comparing clinical outcomes between physicians using the explainable AI system versus standard care would provide definitive evidence of clinical utility and cost-effectiveness necessary for healthcare system adoption and reimbursement decisions. Sixth, investigating federated learning architectures that enable multi-institutional model training without centralizing sensitive patient data would address privacy concerns and regulatory barriers while leveraging larger and more diverse datasets to improve model robustness. Finally, developing calibration and uncertainty quantification methods that provide confidence intervals for individual predictions would support appropriate clinical interpretation by indicating when the model operates within versus outside its training distribution, enabling physicians to appropriately weight AI recommendations alongside other clinical evidence.

Author Contributions: Conceptualization, C.R. and R.S.R.; methodology, C.R. and C.N.S.K.; software and implementation, V.D.; validation and formal analysis, V.D. and K.A.; data curation and preprocessing, C.N.S.K. and K.A.; writing—original draft preparation, C.R.; writing—review and editing, R.S.R. and C.N.S.K.; visualization and explainability analysis, V.D.; supervision and project administration, C.R. and R.S.R. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors gratefully acknowledge the institutional support from CVR College of Engineering, BVRIT HYDERABAD College of Engineering for Women, Anurag Engineering College, Geethanjali College of Engineering and Technology, and Malla Reddy College of Engineering. We thank The Cancer Genome Atlas Research Network and the MIMIC-III database contributors for making genomic and clinical datasets publicly available. We acknowledge computational resources provided by institutional high-performance computing facilities that enabled model training and validation experiments.

Conflicts of Interest: The authors declare no conflicts of interest. This research received no external funding, and all authors contributed independently without commercial interests or financial relationships that could inappropriately influence the work.

References

1. Wainberg, M.; Merico, D.; Delong, A.; Frey, B.J. Deep learning in biomedicine. Nat. Biotechnol. 2018, 36, 829-838.
2. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; Sundberg, P.; Yee, H.; Zhang, K.; Zhang, Y.; Flores, G.; Duggan, G.E.; Irvine, J.; Le, Q.; Litsch, K.; Mossin, A.; Tansuwan, J.; Wang, D.; Wexler, J.; Wilson, J.; Ludwig, D.; Volchenboum, S.L.; Chou, K.; Pearson, B.; Madabushi, S.; Shah, N.H.; Butte, A.J.; Howell, M.D.; Cui, C.; Corrado, G.S.; Dean, J. Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. 2018, 1, 18.
3. Liang, H.; Tsui, B.Y.; Ni, H.; Valentim, C.C.; Baxter, S.L.; Liu, G.; Cai, W.; Kermany, D.S.; Sun, X.; Chen, J.; He, L.; Zhu, J.; Tian, P.; Shao, H.; Zheng, L.; Hou, R.; Hewett, S.; Li, G.; Liang, P.; Zang, X.; Zhang, Z.; Pan, L.; Cai, H.; Ling, R.; Li, S.; Chen, Y.; Lee, C.S.; Lee, A.Y.; Oakley, J.D.; Faleer, A.; Wang, K.; Zhang, H.; Xia, H. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat. Med. 2019, 25, 433-438.
4. Shameer, K.; Johnson, K.W.; Glicksberg, B.S.; Dudley, J.T.; Sengupta, P.P. Machine learning in cardiovascular medicine: are we there yet? Heart 2018, 104, 1156-1164.
5. Holzinger, A.; Dehmer, M.; Emmert-Streib, F.; Cucchiara, R.; Augenstein, I.; Del Ser, J.; Samek, W.; Jurisica, I.; Díaz-Rodríguez, N. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf. Fusion 2021, 79, 263-278.
6. Huang, Z.; Zhu, X.; Ding, M.; Zhang, X. Medical multimodal fusion with deep neural networks. Brief. Bioinform. 2020, 21, 440-449.
7. Cheerla, A.; Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 2019, 35, i446-i454.
8. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal biomedical AI. Nat. Med. 2022, 28, 1773-1784.
9. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? arXiv 2017, arXiv:1712.09923.
10. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems; Curran Associates: Red Hook, NY, USA, 2017; Volume 30, pp. 4765-4774.
11. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13-17 August 2016; pp. 1135-1144.
12. Samek, W.; Wiegand, T.; Müller, K.R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv 2017, arXiv:1708.08296.
13. Steyerberg, E.W.; Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur. Heart J. 2014, 35, 1925-1931.
14. Sendak, M.P.; Gao, M.; Brajer, N.; Balu, S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit. Med. 2020, 3, 41.
15. Vickers, A.J.; Elkin, E.B. Decision curve analysis: a novel method for evaluating prediction models. Med. Decis. Making 2006, 26, 565-574.
16. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015, 350, g7594.