RJNMSR
Review Journal of Neurological & Medical Sciences Review

Volume 4, Issue 3, 2026

DEEP LEARNING WITH ENRICHED PATHOGENICITY SCORES FOR HBB VARIANT CLASSIFICATION IN β-THALASSEMIA

Raazia Sosan Waseem[*1], Muhammad Hussain Habib[2]

[*1]*DHA Suffa University, Karachi, Pakistan*
[2]*Salim Habib University, Karachi, Pakistan*

[*1]raazia.sosan@dsu.edu.pk

Corresponding Author: *
Raazia Sosan Waseem
DOI: http://doi.org/10.5281/zenodo.19365329

Received	Accepted	Published
01 February 2026	17 March 2026	31 March 2026

ABSTRACT

The monogenic disorder β-thalassemia is a common genetic condition caused by pathogenic variations in the HBB gene, and its accurate computational classification is difficult due to class imbalance, incomplete annotations, and individual predictor limitations. This work proposes a deep learning framework for HBB gene variant pathogenicity prediction, which is interpretable, using a combination of REVEL, AlphaMissense, and existing scoring, while using explainable AI for evaluating feature contributions. The dataset of 1585 HBB gene single nucleotide variations was collected from ClinVar, and functional scores were obtained using the myvariant.info API. Five models were implemented using stratified splitting, class imbalance was addressed using SMOTE and class weighting in loss functions. The class weighting deep learning model performed best, with a high ROC-AUC of 0.9483 and PR-AUC of 0.7912, marginally higher than RF and XGB models. Feature importance analysis revealed REVEL as the most dominant predictor, while AlphaMissense contributed significantly in second place. SHAP analysis revealed that REVEL contributes a global predictive value, while AlphaMissense contributes a context-dependent structural value, especially in terms of protein stability. Traditional scoring, such as SIFT, contributed minimally, indicating its redundancy in contributing to the model. This work concludes that the predictive capacity is dependent on the quality of the features and not the architecture, with the combination of sequence-based (REVEL) and structure-based predictors (AlphaMissense) providing complementary and biologically relevant signals. Furthermore, the high recall rates for pathogenic variants and the interpretability through the application of the SHAP analysis also point to the potential usefulness of the model in prioritizing variants of uncertain significance, enhancing diagnostic accuracy, and making informed decisions in the context of β-thalassemia screening, genetic counselling, and therapeutic interventions.

Keywords: *HBB gene; β-thalassemia; pathogenicity; REVEL; AlphaMissense; CADD; SIFT; PolyPhen; ClinVar; gene therapy.*

Introduction

One of the most clinically relevant monogenic disorders in the world is beta-thalassemia, caused by mutations in the HBB gene on chromosome 11p15.4 [1]. It is characterized by decreased or absent synthesis of the beta-globin chains of hemoglobin, resulting in haemolytic anaemia of varying degrees depending on the specific genotype [2]. The impact of beta-thalassemia is particularly high in South Asia, the Middle East, and sub-Saharan Africa, where carrier rates of the condition have been reported to be higher than

five percent of the general population [3]. With the advent of next-generation sequencing technology, thousands of variants in the HBB gene have been recorded in databases such as ClinVar [4] and HbVar [5]. However, a large proportion of these variants have uncertain clinical implications, which directly hampers precise molecular diagnostic techniques and drug development.

The classification of genetic variants into pathogenic and benign types is a crucial step in clinical genomics [6]. Traditional computational methods like SIFT [7], PolyPhen-2 [8], and CADD [9] have been commonly used as predictors of genetic variants in diverse ways by considering various biological signals. Though these methods are useful, they have shown different performance results when tested on gene-specific data sets with class imbalance problems, as observed in the case of the HBB gene in ClinVar [10]. The Combined Annotation-Dependent Depletion (CADD) tool uses more than sixty different annotations and has shown better performance in comparing different predictors in genome-wide deleteriousness predictions [9]. Phylogenetic conservation score predictors like phyloP [11] are useful in understanding the functional importance of genetic variants in highly conserved gene regions like HBB.

Ensemble Meta Predictors represent the latest and most potent variant effect prediction tools. Indeed, a tool named REVEL[12], which combines thirteen individual variant effect prediction tools, such as SIFT, PolyPhen-2, MutPred, VEST, etc., was found to be superior to its individual components in a series of benchmark studies and has since been integrated into clinical guidelines for variant classification by ClinGen [13]. Recently, a new tool named AlphaMissense[14], developed by Google DeepMind, utilized a novel method involving structural embeddings from AlphaFold2 to make predictions on a proteome-scale level, achieving over 90% sensitivity and specificity in a series of benchmark studies. Indeed, other studies have validated the versatility of AlphaMissense in different gene contexts, while also revealing its shortcomings in regions with disordered structures where AlphaFold2 has a low confidence level [15, 16].

The development of machine and deep learning approaches for variant pathogenicity prediction has also seen significant acceleration in recent years [17]. Sundaram et al. have shown the promise of sequence-level variant features for genome-wide pathogenicity inference by training a deep neural network on population-scale variant data [18]. Gene-specific approaches have also been developed for clinically significant disease genes, acknowledging the different genomic and evolutionary contexts of different genes that may warrant different modeling strategies [19]. For haemoglobin disorders in particular, Waseem and Habib [20] have very recently developed a machine learning framework for HBB variant classification based on ClinVar data, which attained a ROC AUC of 0.81 for the three-feature model of variant type, GC content, and phyloP conservation score using the XGBoost algorithm. The authors of this study specifically highlighted the importance of including CADD scores, SIFT scores, and PolyPhen-2 scores in machine learning models for variant classification for this dataset, and the potential of deep learning methods for enhancing variant classification performance.

However, there are some important gaps in the existing literature. For instance, there is a lack of evaluation of the performance of next-generation predictors, like REVEL and Alpha Missense, in gene-specific pathogenicity classification of HBB variants. Another gap in the existing literature lies in the fact that most existing workflows heavily depend on large annotation databases like dbNSFP, which can reach sizes of over 30 GB. Additionally, there is a lack of comparison of the performance of deep learning and classical machine learning techniques, especially in small, imbalanced gene-specific datasets. While some general machine learning frameworks for pathogenic variant prediction have been proposed [37, 38], they have not been applied to HBB-specific datasets with enriched feature sets incorporating next-generation predictors. Most existing works in the area have not shown interpretable results, which can be important in clinical practice. Another gap in the existing

literature lies in the fact that there is a lack of comparison of sequence-based and structure-based predictors, which can be important in practice because of their biological relevance. The impact of class imbalance handling strategies has not been sufficiently explored. Additionally, there is a lack of rigorous analysis of the importance of individual predictors, which can be important in practice. The performance of pathogenicity predictors in different genes has not been sufficiently explored, even though it varies across genes.

The current research aims to bridge some of the gaps in the existing literature by proposing an expanded nine-feature set comprising six functional scores (CADD, SIFT, PolyPhen-2, phyloP, REVEL, AlphaMissense) and three engineered features (transition/transversion type, GC content, and CDS location). The expanded nine-feature set is developed by programmatically accessing the relevant data via the 'myvariant.info' REST API [21] and does not require any large annotation databases. The paper also includes a comparative evaluation of five machine learning and deep learning models on a dataset of 1585 HBB SNVs collected from ClinVar. The paper provides insights into the performance of the models on a small and imbalanced dataset for a gene of interest. The proposed deep learning model is shown to have a ROC AUC of 0.9483 on the dataset of 1585 SNVs and is a significant improvement over the existing benchmarks. The feature importance analysis also shows that the REVEL and AlphaMissense scores are the most important features for predicting the functional impact of SNVs. The SHAP analysis is also conducted to improve the interpretability of the results and assess the potential clinical applicability of the proposed method. The results also indicate that the REVEL and AlphaMissense scores are important for predicting the functional impact of SNVs and have global and context-dependent predictive value.

Materials And Methods

Study Design

The current study is a computational investigation with a binary classification scheme for HBB SNVs. It consists of four stages: data curation using ClinVar, functional score annotation using public APIs, model training with the handling of class imbalance, and comparative evaluation. No human participants were included in the investigation. No experimental procedures were carried out. No ethical committee approval is required. A detailed workflow diagram is presented in Figure 1.

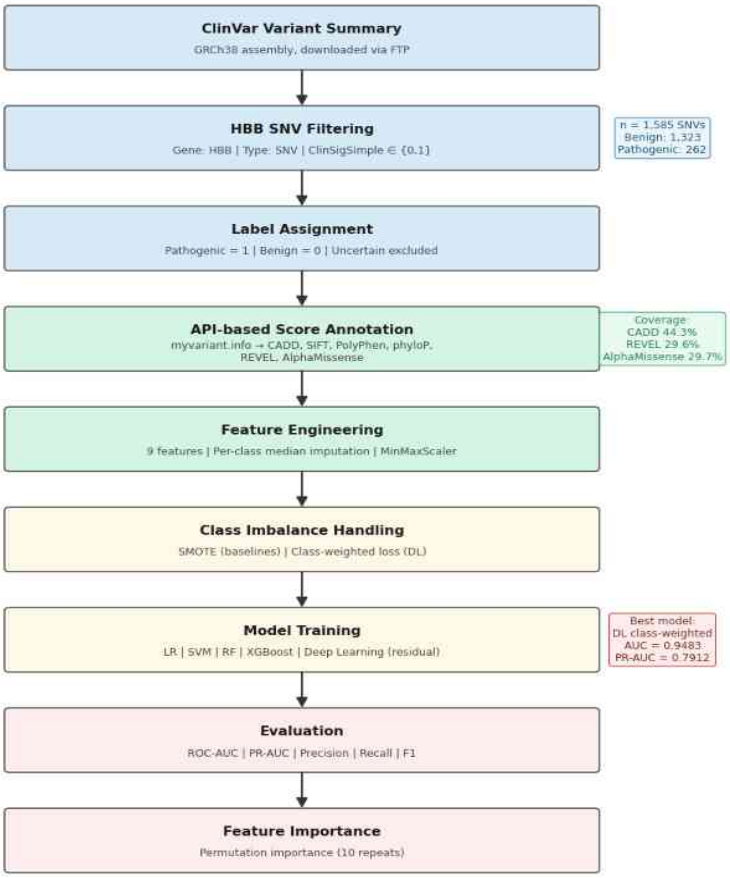

Figure 1: Workflow Diagram

Dataset Construction.

Data on variants was collected using the ClinVar Variant Summary File, which is based on the GRCh38 assembly and was downloaded in March 2026 [4]. Data was filtered to include only those variants involving the HBB gene and single nucleotide variants with unambiguous clinical labels using the ClinSigSimple field. It is assigned the value 1 if the variant is pathogenic/likely pathogenic and 0 if it is benign/likely benign. Variants with uncertain significance, conflicting, and other designations were excluded. Allele data were collected using the ReferenceAlleleVCF and AlternateAlleleVCF fields. These fields supply standardized alleles in VCF format. The dataset comprises 1,585 variants, with 1,323 being benign and 262 pathogenic.

Feature Annotation

The functional pathogenicity score was accessed through the myvariant.info REST API [21], which provides programmatic access to the dbNSFP annotation database. The rsID queries were performed where available for 1,470 variants, with HGVS genomic notation as the fallback for 115 variants. Seven functional pathogenicity scores were accessed: the CADD phred-scaled score [9], the SIFT score with inverted scale to ensure the same directionality as the other scores [7], the PolyPhen-2 HDIV score [8], the phyloP 100 way vertebrate score [11], the REVEL ensemble score [12], the AlphaMissense score [14], and the GERP++ evolutionary constraint score [22] was also queried but excluded from the final feature set due to zero coverage for HBB variants in dbNSFP. In the case where the database returns multiple values for the same variant, the maximum is used. The two engineered features are the

transition/transversion classification and the GC content of the reference sequence. All features are normalized to the range [0, 1] through MinMaxScaler on the training set only.

Missing Value Strategy

The coverage of the score varies depending on the type: CADD 44.3%, SIFT, PolyPhen, REVEL 29.6%, AlphaMissense 29.7%, phyloP 32.9%. The coverage is naturally lower for the coding region as the intronic and synonymous variants do not have the SIFT and PolyPhen score by design. The missing values are imputed through the median imputation strategy on a class basis.

Class Imbalance Handling

The class imbalance is 5:1 between the benign and the pathogenic variants. The class imbalance is addressed through the Synthetic Minority Oversampling Technique (SMOTE) [23] on the training set only, as well as the class-weighted cross-entropy loss, where the weight is inversely proportional to the class frequency. The class-weighted cross-entropy loss is found to provide superior performance for the deep learning approach**.** The effect on the imbalance is presented in Figure 2.

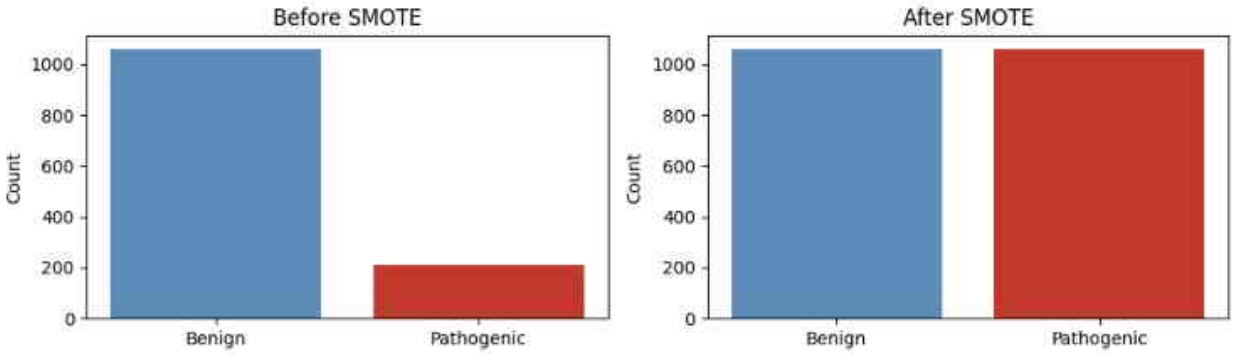

Figure 2: Applying SMOTE to handle the class imbalance

Model Architectures

The performance of five models were examined: Logistic Regression with L2 regularization, Support Vector Machine with a radial basis function kernel, Random Forest with 200 estimators, XGBoost with 200 boost rounds [24], and the proposed deep learning classifier. The deep learning classifier consists of a fully connected residual network with the following structure: input (9) → Layer 1 (64, BatchNorm, ReLU, Dropout 0.3) → Layer 2 (32, BatchNorm, ReLU, Dropout 0.3) → Layer 3 (16, BatchNorm, ReLU, Dropout 0.3) → output (1, Sigmoid), and residual connections from Layer 1 to Layer 3. The model was initialized with the Kaiming normal initializer and used AdamW with learning rate 0.001, weight decay 1e-4, and the learning rate scheduler ReduceLROnPlateau, and early stopping with patience 25.

Evaluation

All models were tested on a stratified test set consisting of 317 variants (20% of the dataset). The main evaluation metrics were ROC AUC and PR AUC, with the optimal classification threshold chosen according to the Youden index. The performance of the Random Forest and deep learning classifier models was evaluated in terms of permutation feature importance, which measures the average decrease in AUC over 10 repetitions for each feature, with the feature being shuffled.

Results

Dataset Characteristics

The final dataset comprised 1,585 HBB SNVs from ClinVar: 1,323 benign/likely benign and 262 pathogenic/likely pathogenic (5:1 ratio). The train-test split yielded 1,268 training and 317 test variants with class proportions preserved by

stratification. After SMOTE, the training set was balanced to 1,058 samples per class (2,116 total).

In the first set of experiments, we implemented a deep learning classifier to our dataset. The baseline deep learning classifier is a residual fully connected network, which is particularly suited for binary classification of HBB variants. The network architecture consists of three hidden layers of different sizes 64-32-16 units, followed by batch normalization, ReLU activation, and dropout (0.3) for better generalization and preventing overfitting. The output layer is followed by a sigmoid activation function for binary classification. The binary cross-entropy loss function is used for balancing the classes in the dataset by giving greater weights to the minority class of pathogenic variants. The proposed network is trained using the AdamW optimizer and a learning rate scheduler with an early stopping criterion. The network is a good trade-off between capacity and robustness for small genomic datasets and is also interpretative for feature attribution methods. The results are presented in Table 1.

Table 1: Performance of Deep Learning Model on Test Set (n = 317)

Class	Precision	Recall	F1-score	Support
Benign	0.98	0.86	0.92	265
Pathogenic	0.57	0.90	0.70	52
Accuracy			**0.87**	317
Macro Avg	0.77	0.88	0.81	317
Weighted Avg	0.91	0.87	0.88	317

The performance of the deep learning model on the test data is satisfactory, with good performance in terms of ROC-AUC (0.9303) and PR-AUC (0.7506) at the optimal threshold of 0.552. The class-wise performance of the model is very high in precision for benign variants (0.98) and high in recall for pathogenic variants (0.90). This is because most of the clinically relevant pathogenic variants are correctly classified by the model, although there are some false positives in the prediction of pathogenic variants (precision of 0.57). The performance of the model is satisfactory in the context of the clinical need for minimizing false negatives in pathogenic variants. The performance of the deep learning model is satisfactory for its use in variant prioritization, reduction of variants of uncertain significance (VUS), and decision-making in genetic screening and diagnosis of β-thalassemia.

This classification performance is further depicted in Figure 2, which shows the confusion matrix, ROC curve, and precision-recall curve. From the confusion matrix, it is evident that the model is able to classify the majority of benign variants (229/265) and pathogenic variants (47/52), with minimal false negatives. From the ROC curve, the model is able to discriminate well between the two variants, with the AUC being approximately 0.93. From the precision-recall curve, the model is able to perform well even in the presence of class imbalance, with the AP being approximately 0.75 [40]. From the optimal threshold selected, it is evident that the model is able to achieve a threshold between sensitivity and specificity, which is approximately 0.55.

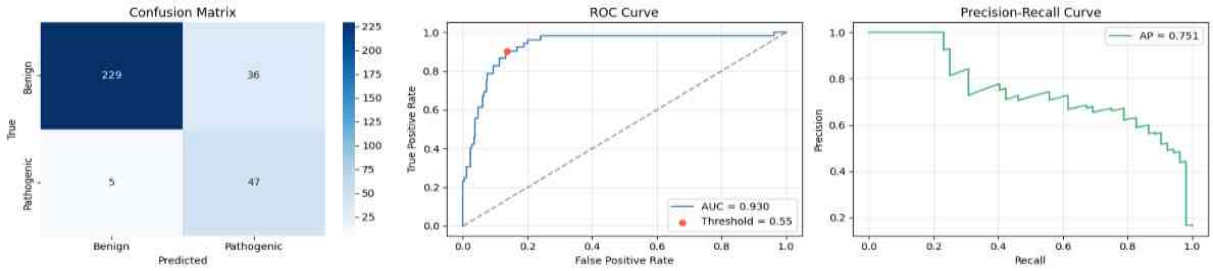

Figure 2. Model performance evaluation on the test set.

The confusion matrix (left), receiver operating characteristic curve (middle), and precision-recall curve (right) for the deep learning model. The confusion matrix illustrates the classification performance of the model at the optimal threshold (Youden's Index ≈ 0.55). The receiver operating characteristic curve confirms the high separability of the classes (AUC ≈ 0.93), while the precision-recall curve (AP ≈ 0.75) emphasizes performance in the case of class imbalance.

Moreover, the permutation feature importance analysis (Table 2 & Figure 3) reveals that REVEL has the highest feature importance, i.e., the highest reduction in AUC by permuting the feature (ΔAUC ≈ 0.039), followed by AlphaMissense (ΔAUC ≈ 0.027). The performance of traditional features such as PolyPhen-2, CADD, and phyloP is moderately affected. SIFT and engineered features have minimal impact. The negligible contribution of the is_transition feature implies that it does not have much independent predictive power. This reveals that the performance of the model is mainly due to the use of advanced predictors such as ensemble predictors and structure-based predictors, which are more likely to capture the relevant information for pathogenicity prediction.

Table 2: Permutation Feature Importance (AUC Decrease)

Feature	Mean AUC Decrease ± SD
REVEL	+0.0388 ± 0.0099
AlphaMissense	+0.0265 ± 0.0071
PolyPhen-2	+0.0185 ± 0.0031
CADD	+0.0180 ± 0.0093
phyloP	+0.0161 ± 0.0055
in_hbb_cds	+0.0060 ± 0.0024
SIFT	+0.0044 ± 0.0021
gc_content	+0.0037 ± 0.0020
is_transition	−0.0000 ± 0.0033

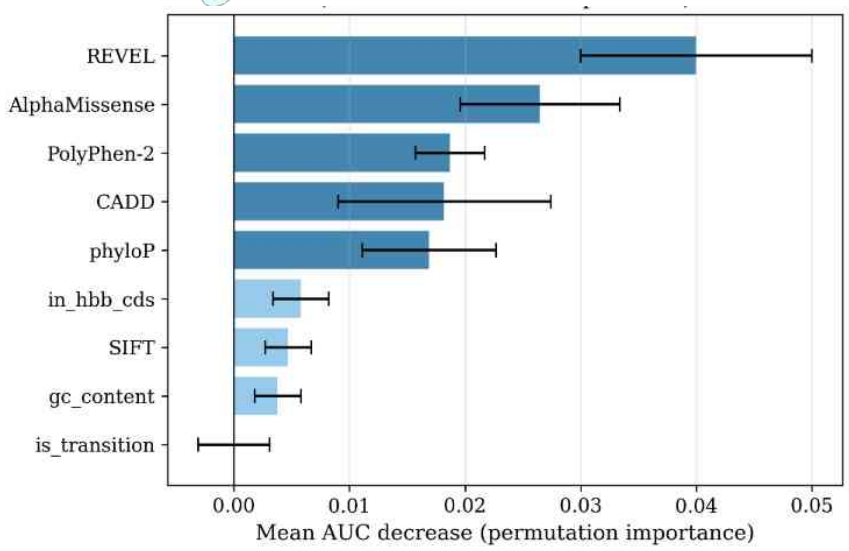

Figure 3: Permutation Figure Importance

In the next set of experiments, a hyperparameter search is performed to assess the performance of different architectural configurations and class imbalance handling techniques. Various

configurations of the deep learning classifier were tested by varying the sizes of the hidden layers, dropout rate, learning rate, and weight decay parameters. Apart from SMOTE-based training, another class imbalance handling technique is class-weighted loss functions. It is observed that varying architectural configurations resulted in a slight difference in performance; however, class-weighted loss achieved better performance compared to SMOTE-based training in terms of ROC-AUC 0.9483 and PR-AUC 0.7869 values. This indicates that class imbalance has a greater impact on performance compared to architectural configurations.

Furthermore, comparison with four different machine learning models is also conducted. The proposed deep learning classifier with a class-weighted loss function showed the best performance in terms of ROC AUC of 0.9483 and PR AUC of 0.7912 (Table 3 & Figure 4). The performance is very close to the performance of the Random Forest and XGBoost models, which showed an AUC of 0.9468 and PR AUC of 0.7923 and an AUC of 0.9440 and PR AUC of 0.7850, respectively, consistent with prior reports of XGBoost achieving competitive performance in variant phenotype prediction tasks [39]. The performance of the SVM and Logistic Regression models is significantly lower than the other models in all metrics, especially for PR AUC of 0.4836 and 0.5196 for SVM and Logistic Regression, respectively. The performance of the basic deep learning classifier trained with SMOTE showed an AUC of 0.9304.

Table 3. Comparative performance of all models on the held-out test set (n=317).

Model	ROC-AUC	PR-AUC	Precision (Path)	Recall (Path)	F1 (Path)	Accuracy
Deep Learning* (proposed)	0.9483	0.7912	–	–	–	–
Random Forest	0.9468	0.7923	0.477	1.000	0.646	0.820
XGBoost	0.9440	0.7850	0.533	0.942	0.681	0.855
Deep Learning (base)	0.9304	0.7507	0.490	0.960	0.650	0.830
SVM	0.8706	0.4836	0.565	0.923	0.701	0.871
Logistic Regression	0.8535	0.5196	0.490	0.904	0.635	0.830

* Class-weighted loss variant. Path = Pathogenic class.

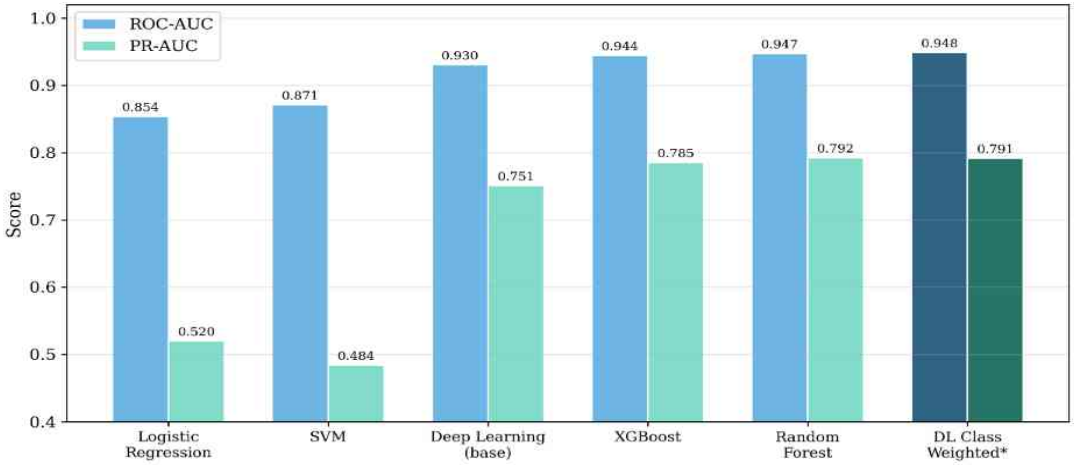

Figure 4: ROC-AUC and PR-AUC comparison of all the models

Moreover, a hybrid deep learning architecture based on feature attention and residual connections has been employed in a validation framework to check whether a more complex architecture results in better performance in terms of predictions. The performance of the hybrid

deep learning model has been reported in Table 4 as 0.9386 ROC-AUC and 0.7452 PR-AUC on the test set. It is observed that there is a slight improvement in performance when compared to the baseline deep learning architecture (0.9304 ROC-AUC) but is outperformed by the class-weighted deep learning architecture (0.9485 ROC-AUC). The performance of each class has been reported as 0.92 recall and 0.55 precision for pathogenic variants, indicating a high recall rate and a low precision rate, thus indicating a high rate of false predictions. The class-wise evaluation revealed high recall for pathogenic variants, at 0.92, but a low precision of 0.55, indicating a high likelihood of false positives. This indicates that while there are marginal benefits in using complex models, they are still not as effective as simpler models when class imbalance handling is appropriately implemented, indicating that in this particular problem, the predictive capability of a model is not necessarily dependent on its complexity but on its features and class imbalance handling strategy. This is a positive aspect in terms of pathogenic variant prediction because there is a low possibility of false negatives, which is critical in genetic screening and diagnostic tests.

Table 4: Comparison of DL Models

Model	ROC-AUC	PR-AUC
Hybrid DL	0.9386	0.7452
DL (base)	0.9304	0.7507
DL (weighted)	**0.9483**	**0.7869**

In the final set of experiments, to better understand the predictions made by deep learning classifier, SHAP analysis has been employed and presented in Figure 5 and 6. The results indicate a strong global influence by REVEL in pathogenicity predictions, followed by AlphaMissense as the second most important predictor in pathogenicity predictions, which has a variable context-dependent influence, particularly in cases of potential structural impact on the HBB protein sequence. Other predictors like PolyPhen-2, CADD, and phyloP have shown moderate influence in pathogenicity predictions, whereas predictors like SIFT and engineered features have shown minimal influence in pathogenicity predictions. From a clinical point of view, these results indicate a strong global influence by REVEL in pathogenicity predictions and thus support the use of integrated sequence and structure predictors in pathogenicity predictions, particularly in cases of VUS.

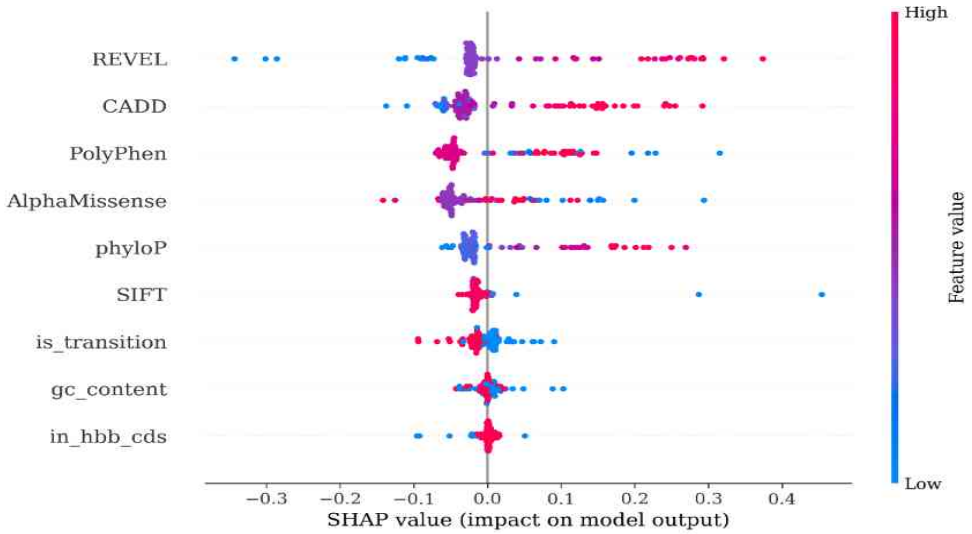

Figure 5. SHAP summary plot showing feature contributions to pathogenicity prediction.

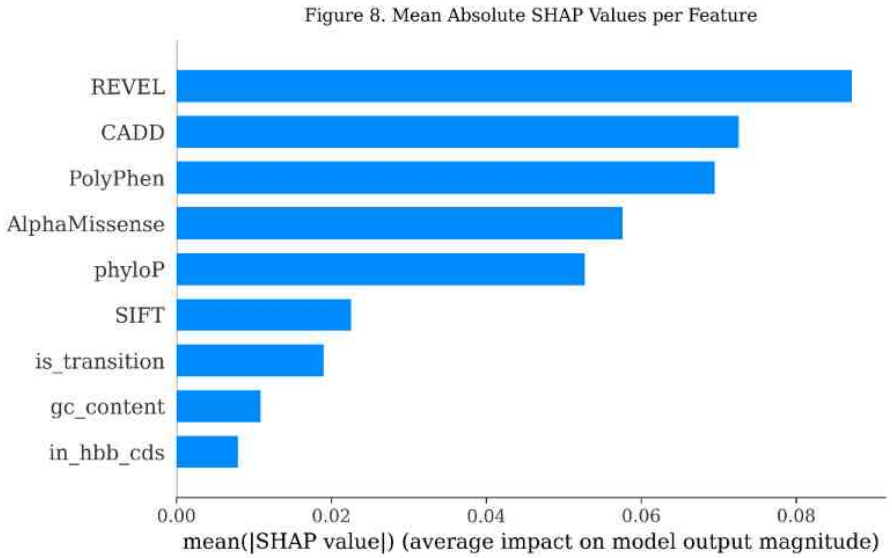

Figure 6. Mean absolute SHAP values indicating global feature importance

Discussion

This study has shown that incorporating next-gen pathogenicity predictors, particularly REVEL and AlphaMissense, significantly enhances HBB variant classification performance compared to earlier studies. The model has a ROC-AUC score of 0.9483, a 17 percentage-point improvement over Waseem and Habib [20], who also worked on the same gene. Significantly, this has been achieved across different model types, implying that performance improvement is mainly linked to feature representation, not model architecture, as also seen in variant classification studies [25, 26].

The prominence of REVEL as a feature aligns with its reputation as a highly accurate ensemble predictor. Ioannidis et al. [12] showed REVEL to be more accurate than individual predictors, and later studies have also shown its robustness in variant classification studies, particularly in clinical settings [13, 27]. The performance of AlphaMissense also supports the integration of structure-based predictors, particularly for well-structured proteins such as HBB, where AlphaFold2-based representation has been shown to be reliable [15, 16]. Overall, these studies point to the synergy between sequence-based and structure-based predictors, allowing for more accurate discrimination of pathogenic variants on the basis of biological meaningful information [18, 28].

Although more complex models, such as hybrid models with attention and residual connections, were also tested, they failed to perform as well as the class-weighted deep learning model. This is consistent with earlier studies showing tree-based and simple models to remain competitive even on small tabular datasets [29]. The performance improvement in this study was mainly linked to class weighting, a more appropriate approach to class imbalance than the generation of synthetic data seen in SMOTE.

From the clinical point of view, the model shows high recall for pathogenic variants, which is important in the clinical setting because false negatives can lead to missed diagnoses. This is the main advantage of the model, making it best suited for the prioritization of variants in the diagnosis of β-thalassemia. It is important to note that the model can be applied in the prioritization of variants in the diagnosis of β-thalassemia because the early identification of pathogenic variants is important in the management of the disease [31, 32]. The accurate prioritisation of pathogenic HBB variants also directly informs the selection of targets for emerging gene editing approaches such as CRISPR-Cas9 [30]. Furthermore, the application of the SHAP-based interpretability is

important because it can be used to resolve variants with uncertain significance. From the methodological point of view, the work is significant because it applied the myvariant.info REST API, which is important because it can replace the annotation tools ANNOVAR[33].

This work also has some limitations. First, the data is limited to 1,585 HBB SNVs with clear ClinVar labels. Variants of uncertain significance were excluded to ensure the quality of the labels. Second, the data on the functional score is not exhaustive, ranging from 29.6 to 44.3 percent. Third, the missing values were handled using the median imputation method per class. Fourth, the data was not validated using an independent set of samples that is not derived from ClinVar.

Conclusion

This study presents an interpretable machine learning framework for pathogenicity classification of HBB gene variants by integrating an enriched set of functional predictors, including REVEL and AlphaMissense, retrieved through the myvariant.info API. The proposed class-weighted deep learning model achieved strong performance (ROC-AUC 0.9483, PR-AUC 0.7912), outperforming baseline approaches and demonstrating that predictive performance is primarily driven by feature quality and appropriate handling of class imbalance rather than model complexity.

Feature importance and SHAP analyses consistently identified REVEL as the most influential predictor, with AlphaMissense providing complementary structural insights. This combination highlights the value of integrating sequence-based and structure-based predictors for improved variant classification. The model's high recall for pathogenic variants is particularly relevant in clinical settings, where minimizing false negatives is critical for accurate diagnosis.

From a clinical perspective, this framework supports the prioritization of variants of uncertain significance, enhances confidence in variant interpretation, and may assist in genetic screening, carrier detection, and therapeutic decision-making in β-thalassemia. Additionally, the use of an API-based annotation pipeline improves reproducibility and accessibility, making the approach suitable for deployment in resource-limited settings.

Overall, this work demonstrates that combining biologically informative features with interpretable machine learning provides a practical and clinically relevant solution for gene-specific variant classification.

Future Work

Future work should focus on improving both the robustness and generalizability of the proposed framework. First, external validation on independent variant datasets beyond ClinVar is necessary to assess model performance across diverse populations and clinical settings. Incorporating population frequency data, such as gnomAD [34, 35], may further improve specificity by reducing false-positive predictions.

Second, expanding the feature space through integration of multi-omics data, including gene expression and epigenomic signals, may provide additional biological context for variant interpretation [36]. The inclusion of regulatory and non-coding variants would also broaden the applicability of the model beyond coding regions.

Third, extending the framework to related haemoglobin genes (e.g., HBA1, HBA2, HBD) and other monogenic disorders would help evaluate its generalizability across gene-specific contexts. Finally, prospective evaluation in clinical workflows and benchmarking against established diagnostic pipelines will be essential to assess real-world utility and clinical impact.

Ethical Approval

This research utilised publicly available, de-identified data from the ClinVar database and the myvariant.info API. No human participants were recruited, no biological samples were collected, and no personal health information was accessed. Ethical approval was therefore not required.

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this manuscript.

Author Contributions

R.S.W. conceptualised the study, designed the computational pipeline, performed data curation, feature engineering, model development, and statistical analysis, and drafted the manuscript.

M.H.H. provided biological and clinical domain expertise for HBB and β-thalassemia, contributed to clinical interpretation of results, and critically reviewed the manuscript for intellectual content. All authors approved the final version for publication.

Acknowledgements

The authors acknowledge the ClinVar database maintained by the National Center for Biotechnology Information (NCBI) and the myvariant.info service for providing free programmatic access to variant annotation data. No funding was received for this study.

REFERENCES

1. Orkin SH, Bauer DE. Emerging genetic therapy for sickle cell disease and thalassemia. Annu Rev Med. 2019;70:257-271.
2. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. Blood. 2010;115(22):4331-4336.
3. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. Bull World Health Organ. 2008;86(6):480-487.
4. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067.
5. Giardine B, Borg J, Viennas E, et al. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. Nucleic Acids Res. 2014;42(D1):D1063-D1069.
6. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-332.
7. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-3814.
8. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249.
9. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D894.
10. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163-2177.
11. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-121.
12. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877-885.
13. Tian Y, Pesaran T, Chamberlin A, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep. 2019;9:12752.
14. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
15. Tordai H, Torres O, Csepi M, Padanyi R, Lukacs GL, Hegedus T. Analysis of AlphaMissense data in different protein groups and structural context. Sci Data. 2024;11:495.

16. Sayers EW, Fogarty ML, Boulger L, et al. Discordance between a deep learning model and clinical-grade variant pathogenicity classification in a rare disease cohort. npj Genom Med. 2025;10:15.
17. Singh PR, Bhatt DL, Creighton CJ, et al. Machine learning approaches for prioritizing pathogenic variants in human disease genes. Genome Med. 2025;17:1124.
18. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-1170.
19. Ahmad RM, Musharraf SM, Al-Amin AQ, et al. A review of genetic variant databases and machine learning tools for predicting pathogenicity. Brief Bioinform. 2024;25(1):bbad479.
20. Waseem RS, Habib MH. Machine learning-based pathogenicity prediction and prioritization of HBB gene variants using ClinVar and population-scale genomic data. Res Med Sci Rev. 2026;4(3):918-928.
21. Xin J, Mark A, Afrasiabi C, et al. High-performance web services for querying gene and variant annotation. Genome Biol. 2016;17:91.
22. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.
23. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357.
24. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785-794.
25. Heo JY, Kim JH. Assessing pathogenicity prediction methods using XGBoost and conservation metrics. BMC Genomics. 2025;26:11787.
26. Danner M, Saez-Rodriguez J, Korbel JO. Predicting pathogenicity of missense variants using machine learning. NAR Genom Bioinform. 2025;7(3):lqaf097.
27. Gunning AC, Fryer V, Fasham J, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet. 2021;58(8):533-541.
28. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91-95.
29. Grinsztajn L, Oyallon E, Varoquaux G. Why tree-based models still outperform deep learning on tabular data. Adv Neural Inf Process Syst. 2022;35:507-520.
30. Frangoul H, Altshuler D, Cappellini MD, et al. CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. N Engl J Med. 2021;384(3):252-260.
31. Cao A, Galanello R. Beta-thalassemia. Genet Med. 2010;12(2):61-76.
32. Taher AT, Weatherall DJ, Cappellini MD. Thalassaemia. Lancet. 2018;391(10116):155-167.
33. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.
34. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-443.
35. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625(7993):92-100.
36. Huang X, Besmer P, Bhatt DL, et al. Epigenomic and transcriptomic profiling of beta-thalassemia pathogenesis. Blood Adv. 2023;7(4):581-594.

37. Bahmane K, Lasserre C, Vandenbon A, et al. PathoPredictor: A machine learning framework for predicting pathogenic missense variants. J Genome Eng. 2026;1(1):3.
38. Zhan H, Zhang Z. Machine learning approaches for variant pathogenicity prediction. arXiv. 2024. arXiv:2406.00164.
39. Xu Z, Li C, Zhou B, et al. Phenotype-aware prioritization of pathogenic variants using XGBoost. Nat Commun. 2023;14:43651.
40. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proc 23rd Int Conf Mach Learn. 2006:233-240.