# Machine Learning Classification of Hemoglobin Beta Gene Mutations

Anja Radomirović

Faculty of Computer Science, University Union,
Belgrade, Serbia

**Abstract.** Mutations in the HBB gene cause severe hemoglobinopathies such as sickle cell disease and beta-thalassemia. Accurate HBB variant classification is crucial for diagnosis but remains challenging. I present a bioinformatics pipeline integrating HGVS parsing, Ensembl annotation, SpliceAI, and BioPython to analyze 1,809 ClinVar variants. Seven models were trained with SMOTE. XGBoost achieved an F1-score of 0.9495 and perfect recall, though ROC-AUC 0.4489 showed discrimination limits. Results highlight ML challenges for single-gene classification and importance of data quality in genomic medicine.

**Keywords:** HBB gene, variant pathogenicity, machine learning, protein encoding, XGBoost, hemoglobinopathies

## 1 Introduction

Existing in silico tools like PolyPhen-2, SIFT, and CADD are widely used for mutation pathogenicity prediction [1], but their different algorithms often produce conflicting results and may lack gene-specific accuracy. To address this, I developed a specialized machine learning model focused on the HBB gene, combining sequence and structural protein features for more precise analysis. By leveraging detailed knowledge of HBB, including exon-intron structure, functional domains, and known pathogenic variants, our approach improves clinical interpretation, supports diagnostics and genetic counseling, and facilitates personalized medicine for hemoglobinopathies.

### 1.1 Motivation and Problem Statement

Hemoglobinopathies represent one of the most common inherited disorders worldwide, affecting approximately 7% of the global population as carriers and resulting in over 300,000 affected births annually [2]. Beta-thalassemia and sickle cell disease, caused by mutations in the HBB gene,

David C. Wyld et al. (Eds): MLCL, CSITA, NMCO, AIRN, CONEDU – 2026
pp. 75-93, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160106

imposes a significant burden on healthcare systems and the patient quality of life, particularly in regions such as the Mediterranean, the Middle East, and Southeast Asia. Despite the clinical importance of accurately classifying HBB variants, the current variant interpretation remains challenging due to several factors.

HBB variant interpretation is challenging due to the wide spectrum of mutation effects, from benign polymorphisms to severe beta-thalassemia, and the frequent conflicting predictions of universal tools like PolyPhen-2, SIFT, and CADD [3]. Many variants are also classified as "Variants of Uncertain Significance" (VUS), limiting clinical utility. A gene-specific machine learning model that integrates sequence, splicing, and protein-level effects can improve classification accuracy and consistency, enhancing molecular diagnosis and genetic counseling.

### 1.2 Research Objectives

The main objectives of this research are:

- Development of a bioinformatics pipeline for translating DNA sequence → mRNA → protein, enabling the identification of mutation effects at all levels of gene expression.
- Implementation and validation of a pre-trained machine learning model for binary classification of mutations (benign/pathogenic) specific to the HBB gene.

### 1.3 Contributions

This work makes several novel contributions to the field of computational variant interpretation:

- Comprehensive bioinformatics pipeline: I developed an end-to-end automated workflow that processes HBB mutations from genomic coordinates through transcription and translation, generating complete mutant protein sequences. The pipeline integrates HGVS notation parsing[4], Ensembl REST API for genomic annotation[5], SpliceAI for splice site effect prediction[6], and BioPython for accurate codon-to-amino acid translation[7].
- Systematic evaluation of machine learning approaches: compared seven different classifiers (XGBoost[8], Random Forest, Gradient Boosting, Neural Network, SVM, Logistic Regression, Naive Bayes) for HBB variant pathogenicity prediction, providing detailed performance analysis across multiple metrics (accuracy, precision, recall, F1-score, ROC-AUC).

- Integrated feature representation: In addition to one-hot encoding of protein sequences, I expanded the feature set with biologically relevant mutation characteristics, including mutation type, premature stop codons, protein length changes, and predicted splice effects [6], creating a comprehensive representation of variant impact.
- Critical analysis of dataset limitations: Through rigorous evaluation, we identified and documented fundamental challenges in HBB variant classification, including severe class imbalance (234 pathogenic vs. 25 benign variants), high-dimensionality issues (57,603 features vs. 207 training samples), and the limitations of synthetic oversampling (SMOTE[9]) in this context.
- Curated HBB mutation benchmark: It is processed and curated 1,809 HBB variants from ClinVar[10], successfully analyzing 1,804 (99.7%) with complete genomic-to-protein translation, providing a valuable resource for future research in gene-specific variant interpretation.

While results revealed significant challenges that prevent immediate clinical deployment, this work establishes a foundation for future improvements through advanced protein language models, expanded datasets, and integration of 3D structural information.

## 2 Related Work

### 2.1 General Computational Methods for Variant Interpretation

Early variant prediction tools used evolutionary conservation: SIFT predicts deleterious substitutions but misses non-conserved regions, PolyPhen-2 [11] adds structural info, and CADD [12] combines 60+ annotations via machine learning, though all lack gene-specific sensitivity.

Advanced methods like PrimateAI [13] and REVEL [14] improve robustness but treat genes uniformly. Splice prediction includes MaxEntScan [15] and SpliceAI [16], which models long-range effects with high accuracy.

### 2.2 Computational Studies of HBB Mutations

HBB mutations are well-characterized clinically, with databases like HbVar [17] cataloging over 1,000 variants. However, computational models targeting HBB are limited. Steinberg et al. [18] and Borg et al. [19] provided clinical and genotype–phenotype insights but no predictive models. Feng et al. [20] applied machine learning to a small set of pathogenic variants without full genomic-to-protein translation, highlighting the need for gene-specific computational frameworks.

### 2.3 How this work differs

This work differs from existing approaches in several key aspects:

1. Gene-specific pipeline: I developed a workflow specifically for HBB, incorporating gene-specific biology such as exon-intron structure, canonical transcript selection, and hemoglobin-specific functional constraints.
2. Full genomic-to-protein translation: The pipeline translates DNA → mRNA → protein for each variant, handling complex mutations (frameshifts, indels, splice-affecting variants) beyond pre-computed annotations.
3. Multi-level feature integration: I combined SpliceAI splice predictions with protein-level consequences (premature stops, length changes, frameshift detection) to create a comprehensive variant representation.

These results show that one-hot encoding and traditional ML classifiers alone are insufficient for accurate HBB variant classification. This work establishes a foundation for future improvements through protein language models (ESM-2), expanded training data from functional assays, and integration of 3D structural information from AlphaFold2 predictions.

## 3 Background

### 3.1 Hemoglobin Structure and Function

Hemoglobin is a tetrameric protein composed of two alpha and two beta chains, each containing a heme group that binds oxygen, allowing one molecule to carry up to four oxygen molecules. The beta-globin chain, encoded by the HBB gene, consists of 146 amino acids and is essential for cooperative oxygen binding, enabling efficient oxygen uptake in the lungs and release in tissues. Even single amino acid substitutions in beta-globin can disrupt hemoglobin stability and function, leading to severe clinical disorders.

### 3.2 Genetic Mutations: Molecular Classification

Translation occurs in the cytoplasm where ribosomes decode mRNA into polypeptide chains using the genetic code, which operates in triplets (codons). Translation begins at the AUG start codon (methionine) and terminates at stop codons (UAG, UAA, UGA)[22, 26]. The triplet nature creates three possible reading frames (0, 1, 2); insertions or deletions not

divisible by three cause frameshift mutations, altering all downstream codons.

Mutations in the HBB gene can be classified by their molecular mechanism and effect on the resulting protein:

Single Nucleotide Variants (SNVs): Substitutions with variable functional impact. *Synonymous mutations* do not alter amino acid sequence (e.g., GCC→GCT both encode alanine) but may affect splicing or mRNA stability. *Missense mutations* change the encoded amino acid—conservative substitutions maintain similar properties (leucine→isoleucine), while non-conservative changes (glutamic acid→valine in HbS) often disrupt protein structure. *Nonsense mutations* create premature stop codons (CAG→TAG), producing truncated, unstable proteins that cause $\beta$-thalassemia[21].

Insertions and Deletions (Indels): In-frame indels (multiples of three nucleotides) add or remove amino acids without shifting the reading frame and may be tolerated in flexible regions. Frameshift mutations (not divisible by three) alter all downstream codons, typically introducing premature stop codons and causing severe $\beta$-thalassemia phenotypes.

## 4 Materials and Methods

### 4.1 Data Source and Dataset Preparation

Mutation data were obtained from the ClinVar, a publicly available, free database maintained by NCBI, containing information on the relationships between genetic variations and human phenotypes, supported by the NIH and in collaboration with the ClinGen initiative. The dataset included:

- Variant ID
- Chromosome (chromosome 11 for the HBB gene)
- Mutation position
- Reference allele
- Alternate allele
- HGVS mutation notation
- Clinical significance
- Mutation type (deletion, insertion, SNV)
- Gene name (HBB)

**Example:** `15534, 11, 5225677, T, G, NC_000011.10:g.5225677T>G, other, no_assertion_criteria_provided, single_nucleotide_variant, HBB`

The reference mRNA sequence NM_000518.5 was also downloaded from NCBI in FASTA format. This mRNA has a coding sequence (CDS) length of 444 bp and 148 amino acids.

**Clinical Classification Distribution** Table 1 shows the distribution of clinical classifications in the dataset.

**Table 1.** Distribution of clinical classifications in the ClinVar HBB dataset ($n$ = 1,809)

| Classification | Count | Percentage |
|---|---|---|
| Likely_benign | 801 | 44.28% |
| Pathogenic | 283 | 15.64% |
| Uncertain_significance | 261 | 14.43% |
| Other | 189 | 10.45% |
| Pathogenic/Likely_pathogenic | 103 | 5.69% |
| Likely_pathogenic | 47 | 2.60% |
| Conflicting_classifications | 45 | 2.49% |
| Pathogenic—Other | 30 | 1.66% |
| Benign/Likely_benign | 25 | 1.38% |
| Miscellaneous | 25 | ¡1% |
| **Total** | **1809** | **100%** |

### 4.2 Mutation Processing Pipeline

The mutation processing pipeline transforms raw genomic coordinates into complete mutant protein sequences through a multi-stage workflow. The pipeline handles diverse mutation types including SNVs, insertions, deletions, and complex indels, while accounting for splice effects and strand orientation.

**Stage 1: Genomic Coordinate Mapping** Each mutation is initially represented in HGVS genomic notation (e.g.,

`NC\_000011.10:g.5225677T>G`

). The Ensembl REST API is queried to map these coordinates to the canonical HBB transcript (ENST00000335295) and determine whether the mutation falls within exonic, intronic, or splice site regions (±2 bp from exon-intron boundaries). For mutations on the minus strand (HBB is on chromosome 11 minus strand), reverse complementation is applied to convert genomic coordinates to the 5'→3' orientation of the reference mRNA.

**Stage 2: Splice Effect Prediction** Mutations located at splice sites or within 50 bp of exon-intron boundaries are evaluated using SpliceAI via the Ensembl VEP REST API. SpliceAI returns four delta scores representing predicted changes in splice acceptor gain/loss and donor gain/loss. A threshold of $\Delta > 0.5$ for any score indicates likely splice disruption. Based on the highest delta score, the pipeline predicts the most probable outcome: exon skipping, cryptic splice site activation, or intron retention. For mutations activating cryptic splice sites without clear resolution, the variant is flagged and excluded from downstream analysis.

**Stage 3: Transcript Reconstruction** For exonic mutations, the reference mRNA sequence (NM_000518.5, 444 bp CDS) is copied and the mutation is applied at the corresponding transcript position. Insertions extend the sequence; deletions remove nucleotides; SNVs replace single bases. For splice-affecting mutations, the predicted splice outcome guides transcript modification: exon skipping removes the affected exon, cryptic site activation adjusts exon boundaries, and intron retention includes the intronic sequence. Mutations in deep intronic regions (¿50 bp from exons) are assigned the reference mRNA unchanged, as they are unlikely to affect the mature transcript.

**Stage 4: Translation to Protein** The mutant mRNA is translated using BioPython's `Seq.translate()` method with the standard genetic code table. Translation begins at the start codon (AUG) and continues until a stop codon (UAA, UAG, UGA) is encountered. The pipeline detects and records premature termination codons (PTCs), frameshift mutations (mRNA length not divisible by 3), and protein length changes. Each translated protein is compared to the reference beta-globin (147 amino acids) to identify specific amino acid substitutions, truncations, or extensions.

**Output and Quality Control** For each processed mutation, the pipeline generates: original genomic coordinates, mutation type classification, location type (exonic/intronic/splice), splice prediction scores and outcomes, mutant mRNA sequence, mutant protein sequence, detected protein alterations (missense/nonsense/frameshift), and processing status flags. Mutations that cannot be reliably processed (e.g., ambiguous cryptic splice activation) are flagged with skip reasons for manual review. This comprehensive output enables downstream feature engineering and model train-

ing while maintaining full traceability from genomic variant to protein consequence.

**HBB Gene Structure and Annotation** Ensembl REST API was used to process the mutated sequences and retrieve data on genomic positions. For the HBB gene, eight transcripts were identified, and the canonical transcript ENST00000335295 was used as the most likely transcript utilized by the gene.

**HBB Gene Structure** The selected transcript contains 3 exons and 2 introns. The gene is located on chromosome 11, positions 5,225,464–5,229,395, with orientation -1 (minus strand).

- **Exons:**
  - Exon 1: 5,226,930 – 5,227,071 (142 bp)
  - Exon 2: 5,226,577 – 5,226,799 (223 bp)
  - Exon 3: 5,225,464 – 5,225,726 (263 bp)
- **Introns:**
  - Intron 1: 5,226,800 – 5,226,929 (130 bp)
  - Intron 2: 5,225,727 – 5,226,576 (850 bp)
- **CDS:** 444 bp

**SpliceAI Analysis** Problems arise with mutations located at the splice site or in its vicinity. These changes can lead to incorrect selection of nucleotides that are included in introns or exons. However, predicting the outcome of such mutations is almost entirely random, which is why I used SpliceAI, an open-source deep learning algorithm that predicts splicing defects caused by DNA variations. Its delta score values help filter variants but can be difficult to interpret precisely, and complex variants are often not correctly handled.

In this work, the SpliceAI tool was accessed via the Ensembl VEP REST API, where for each mutation, data including genomic position, reference, and alternative allele were sent. The server returns an estimate of the mutation's effect on splicing as delta score values, which are then used to determine whether the mutation alters splicing and the most likely outcome. Based on this, the corresponding mutated mRNA is generated.

For some mutations, it is known whether they will affect splicing, but in some cases, the effect may be masked, meaning the exact impact is unknown. For such mutations, no mRNA was assigned, and they were not included in further processing.[23]

**Protein Translation** The next step is translating the mutation into protein. Translation is performed using the standard genetic code table, which also defines stop codons (TAA, TAG, TGA) that signal the end of the protein chain. Each mRNA sequence is first converted into the corresponding BioPython Seq object, and then translated into amino acids, recording any premature stop codons and potential frameshifts if the mRNA length is not divisible by three.

After translation, the mutated protein sequences are compared with the reference protein to identify changes, including truncations, extensions, or individual amino acid substitutions. Each mutation is recorded, including its position and type of change. Translation is applied automatically to all processed mutated mRNA sequences, and results are stored in tabular form, including statistics on translation success, the frequency of premature stop codons, frameshifts, and protein changes.

### 4.3 Protein Encoding

To prepare the data for input into the ML model, a module was designed for the numerical representation of protein sequences using one-hot encoding, a standard approach in bioinformatics. Each amino acid in the sequence is converted into a vector of length 22, which includes 20 standard amino acids, a stop codon, and a position for unknown or padding values.

In addition to the sequence itself, the encoding includes additional mutation features, such as changes in protein length, presence of premature stop codons, protein extensions or truncations, and effects on splicing. The clinical significance of mutations is numerically encoded into four categories, enabling direct use in predictive models.

The generated data are stored in a compressed NPZ format along with metadata, including variant identifiers, chromosome positions, reference and mutated sequences, allowing for easy access and further analysis. The encoder object is also saved, enabling reuse of the same encoding without recalculation.[24]

### 4.4 Classification Models

Seven classification algorithms were implemented: Random Forest, XGBoost, Gradient Boosting, Multilayer Perceptron (MLP), SVM, Logistic Regression, and Naive Bayes. Input data consisted of one-hot encoded protein sequences (22-dimensional binary vectors per amino acid position) augmented with seven biological features (protein length changes,

premature stop codons, truncations, extensions, frameshifts, and splice effect scores). All features were standardized using StandardScaler before training.

Data were split 80/20 into training and test sets. SMOTE (Synthetic Minority Over-Sampling Technique) was applied to the training set, generating synthetic benign samples to balance classes (187 benign, 187 pathogenic). The test set remained unmodified (5 benign, 47 pathogenic) to preserve natural class distribution. Five-fold cross-validation assessed stability and overfitting risk. Models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The best model was selected based on F1-score and saved with scaling parameters for deployment.

### 4.5 Data Balancing and Limitations

SMOTE increased benign samples from 25 to 187 through interpolation in feature space, enabling balanced training. However, synthetic samples may introduce artifacts: overfitting to non-biological feature patterns, overestimated generalization performance, and creation of implausible variant regions. These risks are inherent to synthetic oversampling on small, imbalanced datasets.

### 4.6 Evaluation Metrics

Model performance was assessed using multiple complementary metrics to provide comprehensive evaluation across different aspects of classification quality. Given the severe class imbalance in our dataset, reliance on any single metric would be misleading.

**Confusion Matrix-Based Metrics** Binary classification uses four basic values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). From these, we derive key evaluation metrics:

- **Accuracy:**
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
Measures overall correctness but can be misleading on imbalanced data.
- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
Proportion of predicted positives that are correct.

– **Recall (Sensitivity):**

$$\text{Recall} = \frac{TP}{TP + FN}$$

Proportion of actual positives that are correctly identified.

– **F1-Score:** Harmonic mean of precision and recall, balancing both metrics:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

**ROC-AUC:** Evaluates model discrimination across all thresholds. AUC = 0.5 indicates random prediction, ¡ 0.5 indicates systematic misranking, and 1.0 indicates perfect classification. This metric is critical for imbalanced datasets as it is insensitive to class distribution.

## 5 Results

### 5.1 Mutation Processing and Translation

A total of 1,809 HBB mutations were retrieved from the ClinVar database, of which 1,804 (99.7%) were successfully processed using the Ensembl REST API to map genomic positions. Only 5 mutations could not be analyzed due to the activation of cryptic splice sites, which require the full genomic sequence for precise prediction.

Translation into protein sequences was successful for all 1,804 processed mutations using the BioPython module and the standard genetic code table. Results show that 72.2% of mutations (1,303) produce a protein identical to the reference beta-globin of 147 amino acids, which includes synonymous mutations that do not change the amino acid sequence and neutral intronic variants that do not affect splicing. The remaining 27.8% (501 mutations) lead to changes in protein structure: 324 missense substitutions altering a single amino acid, 82 nonsense mutations with premature stop codons, 66 frameshift mutations shifting the reading frame, and 78 protein extensions beyond normal length. The longest detected mutant protein had an extreme length of 2,618 amino acids (compared to the normal 147), which was set as the maximum length for one-hot encoding of all sequences.

### 5.2 Feature Representation

The feature representation strategy directly impacts model performance. Our approach combines sequence-based encoding with explicit biological features to capture both the mutant protein sequence and functional consequences of mutations.

**One-Hot Encoding of Protein Sequences** Protein sequences were encoded using one-hot representation, where each amino acid at each position is represented by a binary vector. The encoding scheme includes:

- 20 standard amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)
- Stop codon (*)
- Unknown/padding symbol (X)

Each position in the protein sequence is encoded as a 22-dimensional binary vector with exactly one element set to 1. Given the maximum observed protein length of 2,618 amino acids (from extreme frameshift/extension mutations), the one-hot encoding produces $2,618 \times 22 = 57,596$ binary features per sample.

For example, the reference beta-globin sequence begins with `MVHLTPEEK...`. The first position (Methionine) is encoded as $[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$ where the 11th position (corresponding to M) is 1 and all others are 0.

**Feature Standardization** Before model training, all features were standardized using scikit-learn's StandardScaler to have zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation computed from the training set. Standardization is critical for models sensitive to feature scales (SVM, Neural Networks, Logistic Regression) and improves convergence during optimization. Tree-based methods (Random Forest, XGBoost) are invariant to monotonic transformations but were also standardized for consistency.

**Limitations of One-Hot Encoding**

- Does not account for the biochemical similarity of amino acids (treats hydrophobic residues as completely distinct)
- Does not include structural protein features (secondary/tertiary structure)
- Does not distinguish the functional importance of positions (enzyme active site vs. surface residue)

### 5.3 Model Performance

The XGBoost model achieved the best results with an F1-score of 0.9495, accuracy of 90.38%, and a perfect recall of 1.0 (100% of pathogenic mutations correctly identified).

However, the model's performance highlights several problematic aspects. First, the ROC-AUC of 0.4489 is below random level (0.5), indicating poor discrimination ability; ROC-AUC measures the model's capacity to distinguish classes across all thresholds, and a value below 0.5 suggests systematic misranking of predictions. Second, the precision for the Benign class is 0.00, meaning the model failed to correctly identify any benign samples in the test set—all 5 benign samples were misclassified as pathogenic, which is critical for clinical applications where false positives are problematic.

Third, the discrepancy between metrics—high F1-score and accuracy versus low ROC-AUC and zero precision for benign samples—indicates a bias toward the pathogenic class. The model has "learned" to classify almost all mutations as pathogenic, which inflates accuracy due to the dominance of pathogenic samples in the test set ($47/52 = 90\%$), but does not reflect true predictive power. Finally, the cross-validation score shows a CV F1-score of 0.7019 substantially lower than the test F1-score (0.9495), further suggesting overfitting and that test set performance does not represent true generalization.

#### Overall Performance Metrics

**Confusion Matrix Analysis** The confusion matrix shown in Figure 1 demonstrates that XGBoost correctly identified all 47 pathogenic mutations (True Positives = 47, False Negatives = 0), but failed to correctly classify any benign mutations (True Negatives = 0, False Positives = 5).

The model effectively classifies all samples as pathogenic regardless of their actual characteristics. This explains the high accuracy of 90.38% — since the test set is 90.4% pathogenic (47/52), a model that always predicts the majority class automatically achieves approximately 90% accuracy.

**ROC Curve** The ROC curve shown in Figure 2 with an AUC of 0.4489 highlights the problem. A value below 0.5 indicates that the model has worse discriminative power than a random classifier. The ROC curve lies

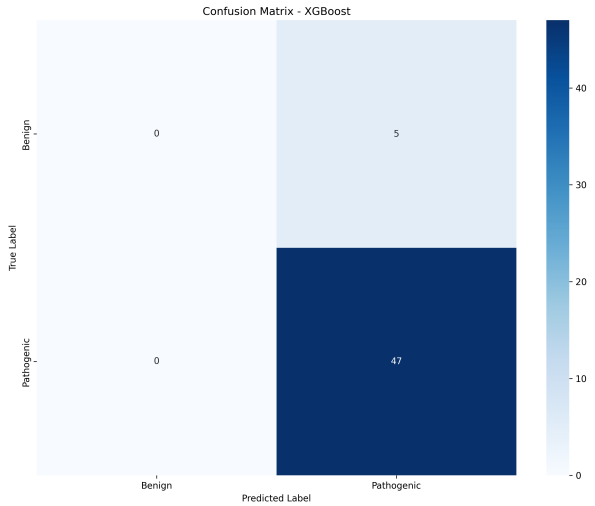


**Fig. 1.** Confusion matrix for the XGBoost model. The model classifies all samples as pathogenic.

below the diagonal, showing that the model systematically misranks predictions across different thresholds.

For comparison, the Neural Network achieved a ROC-AUC of 0.6064 (the best among all models), but with a dramatically lower recall of 0.2979, meaning it misses 70% of pathogenic mutations.

**Comparative Model Analysis** Figure 3 shows a comparative analysis of all seven models. XGBoost achieves the highest F1-score (0.9495) and recall (1.0), but one of the lowest ROC-AUC scores (0.4489). The Neural Network shows the opposite pattern: the best ROC-AUC (0.6064) but poor recall (0.2979).

All models exhibit a *trade-off* between metrics:

- *Tree-based* algorithms (XGBoost, Random Forest, Gradient Boosting) show high recall but low ROC-AUC
- Probabilistic models (Neural Network, Logistic Regression) have better ROC-AUC but miss most pathogenic mutations

No model achieves satisfactory performance across all metrics simultaneously.

**Advantages of Gene-Specific Approach** This pipeline offers several advantages over universal tools:

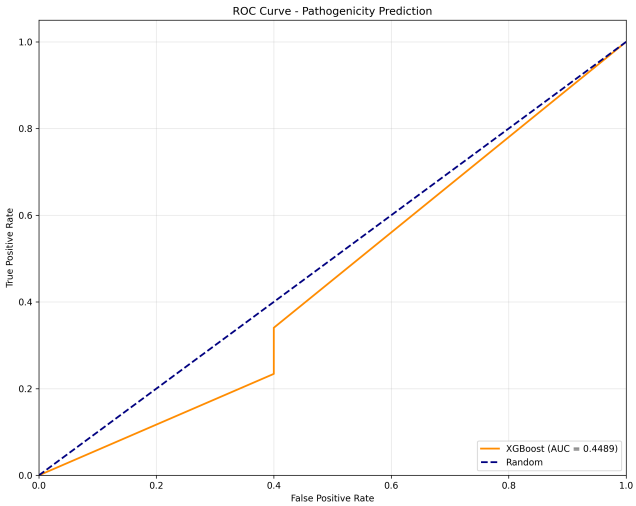


**Fig. 2.** ROC curve for the XGBoost model. AUC = 0.4489, below the random baseline (0.5).

- Complete mutation processing: Full DNA → RNA → protein translation captures complex effects (frameshifts, stop codon readthrough) that annotation-based tools miss.
- Splice integration: Direct incorporation of SpliceAI predictions for variants near exon-intron boundaries provides more accurate transcript consequences than universal tools.
- Transparency: This pipeline provides full traceability from genomic variant to predicted protein sequence to pathogenicity classification, enabling clinical interpretation.
- Extensibility: The modular design allows easy integration of improved encoding methods (protein language models) and additional data sources (functional assays, population databases).

**Current Limitations** However, this approach currently suffers from critical limitations that established tools do not:

- Inadequate training data: approach has only 259 labeled variants.
- Class imbalance: unbalanced datasets 234:25, pathogenic:benign ratio leads to severe bias.
- Feature representation: one-hot encoding discards biochemical information.

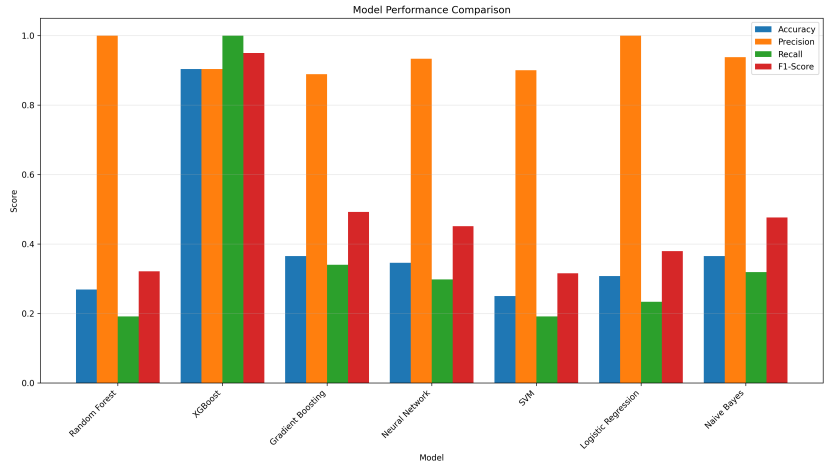


**Fig. 3.** Comparative performance analysis of all models. No model achieves satisfactory performance across all metrics.

Future work will address these limitations through protein language models and expanded training data.

## 6 Discussion

### 6.1 Overview of Achieved Results

In this study, a complete bioinformatics pipeline was developed for analyzing HBB gene mutations and predicting their pathogenicity using machine learning. The pipeline successfully processed 1,809 mutations from the ClinVar database, of which 1,804 mutations were successfully analyzed and prepared for machine learning analysis.

### 6.2 Dataset Quality and Challenges

Analysis of the distribution of clinical classifications revealed a significant imbalance in the data: out of 1,738 valid samples with protein sequences, only 259 mutations (14.9%) had a clear clinical classification as "Pathogenic" (234 samples, 90.3%) or "Benign" (25 samples, 9.7%). The remaining 85.1% of samples were classified as "Uncertain significance" or other ambiguous categories, and had to be excluded from the training set.

This drastic imbalance represents a fundamental challenge in analyzing HBB mutations. The small number of benign samples (only 25)

combined with the strong dominance of pathogenic mutations reflects the nature of the HBB gene—most mutations in this gene lead to clinically significant phenotypes (thalassemia, sickle cell anemia), while benign variants are relatively rare. Additionally, the large number of "uncertain" variants highlights the limitations of current clinical practice in mutation classification.

### 6.3 Limitations and Future Directions

This study revealed fundamental challenges: (1) extreme class imbalance (234:25 pathogenic:benign), (2) high dimensionality (57,603 features vs. 207 samples), (3) one-hot encoding ignoring biochemical properties and 3D structure. Future work requires protein language models (ESM-2, ProtBERT), expanded datasets (gnomAD, HGMD, functional assays), and AlphaFold2 structural predictions.

### 6.4 Proposed Improvements

Expanding the dataset is one of the most important needs and can be achieved by integrating additional databases, such as gnomAD for population frequencies and HGMD for disease-associated mutations. Furthermore, incorporating functional experimental data, such as deep mutational scanning studies, can significantly enrich information on mutation effects.

It is also essential to improve the analysis by implementing protein *folding* verification, i.e., predicting its tertiary structure, alongside investigating the phenotypic consequences of mutations. This approach allows assessment of how specific mutations affect protein stability, conformation, and functionality, improving the biological interpretation of the model and increasing the reliability of predictions in a clinical context.

## 7 Conclusion

This work presents a comprehensive bioinformatics pipeline that I developed for automated HBB gene mutation analysis and pathogenicity prediction using machine learning. I successfully processed 1,804 mutations (99.7% of the ClinVar HBB dataset) through complete genomic-to-protein translation, integrating HGVS parsing, Ensembl annotation, SpliceAI splice prediction, and BioPython translation. I evaluated seven classifiers, with XGBoost achieving an F1-score of 0.9495 and perfect pathogenic recall (1.0).

However, severe class imbalance (234 pathogenic vs. 25 benign) and extreme dimensionality (57,603 features vs. 207 samples) resulted in models lacking discriminative power (ROC-AUC 0.4489, below random baseline). Zero precision for benign variants—all misclassified as pathogenic—demonstrates that high accuracy can mask critical failures in imbalanced datasets. For clinical deployment, accurate identification of both pathogenic and benign variants is essential; a classifier labeling all variants as pathogenic would generate excessive false positives, causing unnecessary patient anxiety, inappropriate interventions, and incorrect risk assessment for families.

To advance toward clinical viability, I propose: (1) dataset expansion through gnomAD, HGMD, and functional assay integration; (2) replacement of one-hot encoding with protein language models (ESM-2, ProtBERT); (3) integration of AlphaFold2 3D structural predictions; (4) ensemble methods combining sequence and experimental data; and (5) rigorous external validation on independent clinical cohorts. Despite current limitations, this work establishes a reproducible framework for gene-specific variant analysis, provides critical insights into challenges of imbalanced genomic ML, and demonstrates that high F1-scores do not guarantee clinical utility—contributing to validation standards for genomic medicine applications.

While substantial refinement is required before clinical deployment, the foundation established here—comprehensive mutation processing, biologically informed feature engineering, and transparent evaluation—represents progress toward reliable computational support for hemoglobinopathy diagnosis and genetic counseling as genomic sequencing becomes increasingly integrated into clinical practice.

## References

1. R. Ghosh, N. Oak, and S. E. Plon, Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines, *Genome Biology*, vol. 18, no. 1, p. 225, 2017.
2. I. Belmokhtar, K.Y. Belmokhtar, S. Lhousni, M. Charif, Z. Sidqi, R. Seddik, M. Choukri, M. Bellaoui, R. Boulouiz, *Carrier frequency and molecular basis of hemoglobinopathies among blood donors in eastern Morocco: Implications for blood donation and genetic diagnosis*, Blood Reviews, 2024. `https://www.sciencedirect.com/science/article/abs/pii/S0009912024001346`
3. Tamana, S., Xenophontos, M., Minaidou, A., Stephanou, C., Harteveld, C.L., Bento, C., Traeger-Synodinos, J., Fylaktou, I., Mohd Yasin, N., Abdul Hamid, F.S., Esa, E., Halim-Fikri, H., Zilfalil, B.A., Kakouri, A.C., Kleanthous, M., Kountouris, P. *Evaluation of in silico predictors on short nucleotide variants in HBA1, HBA2, and HBB associated with haemoglobinopathies.* ClinGen Hemoglobinopathy Variant Curation Expert Panel. *Human Mutation*, 44(11):1985–2003, 2023. `https://pmc.ncbi.nlm.nih.gov/articles/PMC9731569/`.

4. HGVS Nomenclature Resources and Software. Human Genome Variation Society (HGVS). Available at: https://hgvs-nomenclature.org/stable/software/ (Accessed: 17.12.2025).
5. Howe, K. L., Achuthan, P., Allen, J., et al. *Ensembl 2021.* Nucleic Acids Research, 49(D1), D884–D891, 2021. https://doi.org/10.1093/nar/gkaa942
6. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., et al. *Predicting Splicing from Primary Sequence with Deep Learning.* Cell, 176(3):535–548, 2019.
7. Cock, P. J. A., Antao, T., Chang, J. T., et al. *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 25(11):1422–1423, 2009.
8. Chen, T., & Guestrin, C. *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, 2016.
9. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 16:321-357, 2002.
10. Landrum, M. J., Lee, J. M., Benson, M., et al. *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Research, 46(D1):D1062–D1067, 2018.
11. I. A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, *Nature Methods*, vol. 7, pp. 248–249, 2010.
12. M. Kircher, D. M. Witten, P. Jain, et al., A general framework for estimating the relative pathogenicity of human genetic variants, *Nature Genetics*, vol. 46, no. 3, pp. 310–315, 2014.
13. L. Sundaram, H. Gao, S. Padigepati, et al., Predicting the clinical impact of human mutation with deep neural networks, *Nature Genetics*, vol. 50, pp. 1161–1170, 2018.
14. N. M. Ioannidis, V. Rothstein, M. Pejaver, et al., REVEL: An ensemble method for predicting the pathogenicity of rare missense variants, *American Journal of Human Genetics*, vol. 99, no. 4, pp. 877–885, 2016.
15. Yeo, G., & Burge, C. B. *Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.* Journal of Computational Biology, 11(2–3):377–394, 2004. doi:10.1089/1066527041410418. :contentReference[oaicite:0]index=0
16. Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., … Sanders, S. J. *Predicting splicing from primary sequence with deep learning.* Cell, 176(3):535–548.e24, 2019. doi:10.1016/j.cell.2018.12.015. :contentReference[oaicite:1]index=1
17. Giardine, B., Borg, J., Higgs, D. R., Peterson, K., Philipsen, S., Maglott, D., et al. *HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server.* Human Mutation, 28(2):206–213, 2007. https://pubmed.ncbi.nlm.nih.gov/17068035/.
18. Steinberg, M. H., Forget, B. G., Higgs, D. R., & Weatherall, D. J. *Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management.* Cambridge University Press, 2nd edition, 2012.
19. Borg, J., Papadopoulos, P., Georgitsi, M., Gutierrez, L., Grech, G., Fanis, P., et al. *Genotype-phenotype correlations in beta-thalassemia: clinical severity and degree of beta-globin chain reduction.* Haematologica, 94(7): 973–980, 2009. https://pubmed.ncbi.nlm.nih.gov/19498168/.
20. Feng, J., Li, Y., Wang, X., & Zhang, S. *Machine learning approaches for predicting pathogenicity of thalassemia variants.* BMC Bioinformatics, 22:123, 2021. https://pubmed.ncbi.nlm.nih.gov/33658615/.

21. Puglisi, R. *Protein Mutations and Stability, a Link with Disease: The Case Study of Frataxin*. PubMed Central (PMC8962269), PMID: 35203634, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8962269/
22. Clancy, S., & Brown, W. *Translation: DNA to mRNA to Protein*. Nature Education, 1(1):101, 2008.
23. de Sainte Agathe, J. M., et al. *SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation*. Human Genomics, 2023. PubMed Central (PMC9912651). https://pmc.ncbi.nlm.nih.gov/articles/PMC9912651/
24. Baxevanis, A. D., & Ouellette, B. F. F. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 2nd edition. Genome Technology Branch, NHGRI, NIH, Bethesda, MD, USA; Centre for Molecular Medicine and Therapeutics, University of British Columbia, Vancouver, BC, Canada.
25. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12:2825–2830, 2011.
26. Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. *Molecular Biology of the Cell*, 6th edition. New York: Garland Science, 2014.

## Author

I am a second-year undergraduate student in Computer Science at the Faculty of Computer Science (Računarski fakultet), University Union, and a first-year undergraduate student in Computer and Applied Physics at the Faculty of Physics, University of Belgrade. My academic interests focus on computational modeling, machine learning, and their applications in biological and physical systems. I have volunteered twice in projects within the European Solidarity Corps and participated in international climate action initiatives, including the ECO-GEN Youth Conference for COP29.