

MACHINE LEARNING-BASED PATHOGENICITY PREDICTION AND
PRIORITIZATION OF HBB GENE VARIANTS USING CLINVAR AND
POPULATION-SCALE GENOMIC DATARaazia Sosan Waseem^{*1}, Muhammad Hussain Habib²^{*1}DHA Suffa University²Salim Habib University^{*1}raazia.sosan@dsu.edu.pkDOI: <https://doi.org/10.5281/zenodo.21528052>**Keywords**

Genetic Variation; Pathogenicity Prediction; Machine Learning; XGBoost; β -Thalassemia; HBB Gene; Conservation Analysis; Precision Medicine

Article History

Received: 30 December 2025

Accepted: 01 February 2026

Published: 13 February 2026

Copyright @Author**Corresponding Author: *****Raazia Sosan Waseem****Abstract****Background:**

The advent of advanced techniques in genomic sequencing has led to the discovery of several genetic variants, where most of them are still uncharacterized. The determination of the pathogenicity of these genetic variants is crucial in understanding the genetic basis of diseases such as β -thalassemia, caused by mutations in the HBB gene.

Objective:

The study aims to develop a machine learning-based model for pathogenicity prediction and prioritizing novel variations in the HBB gene.

Methodology:

A dataset of 7,819 ClinVar variants, was compiled and labelled as benign or pathogenic. Machine learning models such as Random Forest and XGBoost were implemented and evaluated using techniques for handling class imbalance. The models were evaluated using metrics such as accuracy, ROC-AUC, and precision-recall AUC (PRAUC). The model that outperformed was then applied to the dataset of 36,099 variants in the HBB gene region.

Result:

The XGBoost model was better than the Random Forest model in the prediction of the pathogenicity of the genetic variants. The ROC-AUC and PRAUC were 0.81 and 0.61, respectively. The model showed better precision in the prediction of the pathogenicity of the genetic variants, improving from 0.35 to 0.49 for pathogenic variations. XGBoost was further applied to the dataset of the HBB gene and showed a right-skewed distribution of the probabilities. The model showed that the highly confident predictions (>0.97) were in the highly conserved regions (>7).

Conclusion:

The proposed model is effective in the prediction of the pathogenicity of the genetic variants and the prioritization of novel variations in the HBB gene. The model is applicable in the future in the field of genomic medicine and in the study of diseases.

INTRODUCTION

With the advent of next-generation sequencing technologies, there has been a dramatic increase in the discovery of genetic variants in the human genome. Despite this, the biggest challenge in

genomic medicine today is the interpretation of genetic variants, especially those for which there is no clear clinical significance. Several databases, like ClinVar[1], [2], have been developed to provide detailed information on

the pathogenicity of genetic variants[3]. Despite this, a considerable proportion of genetic variants have either unknown pathogenicity or lack sufficient evidence for interpretation [4].

Accurate prediction of the pathogenicity of genetic variants is essential for understanding the underlying mechanisms of genetic diseases. In this regard, genetic mutations in the β -globin gene, referred to as HBB, cause β -thalassemia, a genetic disorder of great public health significance[5], [6]. Despite extensive research on this gene, many genetic variants within this locus have unknown pathogenicity. Therefore, there is a need for the development of computational methods for prioritizing pathogenic variants.

Machine learning methods have been recognized as a useful tool for the interpretation of genetic variants [7]. These methods have been shown to integrate multiple genomic features, including context, evolution, and function[8]. Among machine learning methods, tree ensemble methods[9], including random forests[10] and extreme gradient boosting[11], [12], have been recognized for their excellent performance in pathogenicity prediction. Boosting methods have been demonstrated to have superior performance compared to traditional classifiers[13].

Additionally, evolutionary conservation features, such as PhyloP[14] and phastCons[15] scores, are also extremely useful, as it has been demonstrated that mutations occurring in highly conserved regions of the genome are more likely to be functionally deleterious [16]. These features are commonly used in modern approaches for predicting pathogenicity and have been shown to improve the performance of such models [17].

Although many tools have been developed for predicting variant pathogenicity[3], [18], many of these tools are focused on genome-wide predictions and do not have direct applications for individual genes[4], [19]. In addition, translating the results of such predictions into useful information for understanding the clinical significance of a variant, such as prioritizing variants occurring in disease-associated genes, remains a challenge[1], [20]. In this study, we present a novel framework for predicting variant pathogenicity using a machine learning-based model trained on ClinVar

variants and apply it to predict the pathogenicity of novel candidate variants occurring in the HBB gene[16], [21].

Materials and Methods

Study Design

This research aims to create a comprehensive framework to forecast the pathogenicity of genetic variants using machine learning and to discover novel variants within the HBB gene. The process involved several steps, from data collection to application.

This study is different from most experimental designs, where the emphasis is on using existing genomic data to create a framework for prediction. In this context, the emphasis is on generalization, where a model is trained on existing variants and then used to predict variants without clear characterization.

Participants

No participants were involved in this study. All the processes were conducted using existing genomic data. The data used in this study were obtained from the ClinVar database, a comprehensive collection of clinically annotated genetic variants and interpretations of their clinical significance[2], [22].

To maintain the quality of the data, only those variants with clearly defined clinical labels, i.e., benign, or pathogenic, were considered. This filtering is an essential step, especially while dealing with supervised learning models, where incorrect information can affect the performance of the model significantly.

Data Collection

The initial dataset for training the model consisted of 7,819 variants, collected from the ClinVar database (GRCh38 assembly) [2], [22], out of which 6,587 variants were benign, while 1,232 variants were pathogenic. For each variant, their genomic coordinates, reference alleles, and alternate alleles were collected.

To check the applicability of the model, another dataset comprising 36,099 variants located within the HBB gene region on chromosome 11, collected from population-scale genomic data, was considered. This dataset is considered an actual scenario where most of the variants are not annotated.

The two different datasets, ClinVar and the unannotated variants, allow us to check the capability of the model to perform variant

prioritization within clinically relevant genomic regions.

The datasets used for the entire process are presented in Table 1.

Table 1. Dataset Summary

Dataset	Total Variants	Benign	Pathogenic	Source
ClinVar (Training)	7,819	6,587	1,232	ClinVar
HBB Variants	36,099	Unknown	Unknown	Population dataset

Feature Extraction

In feature engineering, biologically relevant signals associated with variant pathogenicity were extracted. Three major features were considered for the model. These features included variant type, GC content, and evolutionary conservation. These features were chosen for their relevance to variant pathogenicity, as discussed in the literature. The variant type feature is defined as either a single nucleotide polymorphism or an indel, reflecting the structural difference in genomic variants, which might influence variant pathogenicity. The GC content feature is defined as the GC content within a ± 10 base pair window around the variant. This feature has been associated with

mutation rates and genomic stability. Evolutionary conservation is defined as the PhyloP100way score, which measures conservation at the nucleotide level across multiple species. A high PhyloP100way score is associated with high evolutionary conservation, indicating that mutations in these regions are likely to be functionally deleterious. This feature has been widely accepted as a strong predictor of variant pathogenicity in previous studies [23], [24]

In addition, the combination of sequence and evolutionary features enables the model to learn relevant biological context.

A detailed explanation of the extracted feature is outlined in Table 2.

Table 2. Feature Description

Feature	Description	Type
Variant Type	SNP or Indel	Categorical
GC Content	GC proportion in ± 10 bp window	Numerical
Conservation	PhyloP100way score	Numerical

Statistical Analysis

Model Development

Two supervised machine learning algorithms were used to classify variants as benign or pathogenic: Random Forest (RF)[10] and Extreme Gradient Boosting (XGBoost) [12],[25]. RF builds multiple decision trees using bootstrap aggregation techniques to enhance model robustness and accuracy. On the other hand, XGBoost is a gradient boosting algorithm that optimizes model accuracy by correcting errors in successive iterations.

The data set was found to be imbalanced because benign variants were significantly more abundant than pathogenic variants. Synthetic Minority Over-sampling Technique (SMOTE)

[26], was used to address this imbalance in the training data set. SMOTE generates artificial data points belonging to the minority class to enhance model sensitivity to pathogenic variants and avoid bias towards benign variants[26].

Model Evaluation

The data set was divided into training and testing data sets using an 80:20 ratio to test model accuracy on unseen data. Various model evaluation metrics were used to test model accuracy: accuracy, precision, recall, F1 score, ROC-AUC score, and PR-AUC score. In this case, due to data imbalance, PR-AUC was used as a major evaluation metric because it provides a more reliable assessment of model accuracy in detecting minority class variants than ROC-

AUC score in machine learning problems[26], [27].

Application to HBB Variants

After model evaluation using multiple metrics, the best model (XGBoost) was used to classify HBB variants and predict probabilities of pathogenic variants. The model was able to rank variants in descending order of their probabilities. This ranking is useful in prioritizing variants in genomic studies. In addition, a probability threshold was set to select variants with probabilities above 0.9. This approach is useful in reducing the complexity associated with genomic data and variants.

Ethical Approval

This research used publicly available data obtained from ClinVar database. The data was de-identified and no personal information was used in this research. As a result, no ethical approval.

Results

Model Performance Comparison

The performance of the Random Forest (RF) model and the XGBoost (XGB) model was compared by using different classification metrics on the test dataset. The results show that the XGBoost model performed better than the Random Forest model according to all the classification metrics. In particular, the XGBoost model had an ROC-AUC score of 0.81, while the Random Forest model had an ROC-AUC score of only 0.77. This shows that the XGBoost model is better at distinguishing between pathogenic and benign variants. In addition, the XGBoost model performed better than the Random Forest model regarding the imbalanced dataset, with an PR-AUC score of 0.61, while the Random Forest model had an PR-AUC score of only 0.52. A detailed comparison of model performance is presented in Table 3.

Table 3: Model Performance Comparison

Metric	Random Forest	XGBoost
Accuracy	0.76	0.84
ROC-AUC	0.77	0.81
PR-AUC	0.52	0.61
Precision (Pathogenic)	0.35	0.49
Recall (Pathogenic)	0.62	0.65
F1-score (Pathogenic)	0.44	0.56

Class-wise Performance

A detailed class-wise performance analysis showed that the XGBoost model had a significant impact on the prediction of pathogenic variants. The precision of the pathogenic class increased from 0.35 using RF to 0.49 using XGB. This shows that the number of false positive predictions is reduced using the XGBoost model. Moreover, the recall of the pathogenic class increased from 0.62 to 0.65 using the XGBoost model. This resulted in an

improvement in the F1-score from 0.44 to 0.56 using the RF and XGBoost models, respectively. The performance of the XGBoost and RF models in the prediction of the benign class also showed improvement, with the XGBoost model having better recall and F1-score compared to the RF model. This shows that the XGBoost model provides better and more balanced classification results compared to the RF model. A detailed classification report of the XGBoost model is provided.

Table 4: Classification Report for XGBoost

Class	Precision	Recall	F1-score	Support
Benign (0)	0.93	0.87	0.90	1318
Pathogenic (1)	0.49	0.65	0.56	246
Accuracy			0.84	1564

Receiver Operating Characteristic (ROC) Analysis

The receiver operating characteristic curves for both models are shown in Figure 1. As illustrated, the XGBoost model always maintains a higher true positive rate for a broad range of false positive rates compared to the Random Forest model.

The improvement in the ROC-AUC for the XGBoost model implies a better capacity for distinguishing between benign and pathogenic variants. The gap between the two curves further confirms the superior discriminative ability of the boosting model.

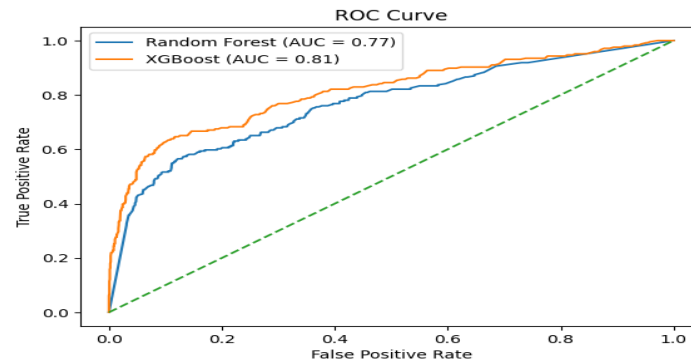


Figure 1: Receiver Operating Characteristic (ROC) curves comparing Random Forest and XGBoost models. The XGBoost model demonstrates improved discriminative performance with a higher AUC (0.81 vs 0.77).

Precision-Recall Analysis

The Precision-Recall curves for both models are presented in Figure 2. The precision values in the XGBoost model are higher compared to those in the RF model for most recall values, reflecting its ability to identify pathogenic variants under class imbalance conditions.

The significant improvement in Precision-Recall-AUC (PR-AUC) from 0.52 in RF to 0.61 in XGB suggests that XGBoost can identify fewer false positives while retaining sensitivity to pathogenic variants in genomic data, where they are known to be rare.

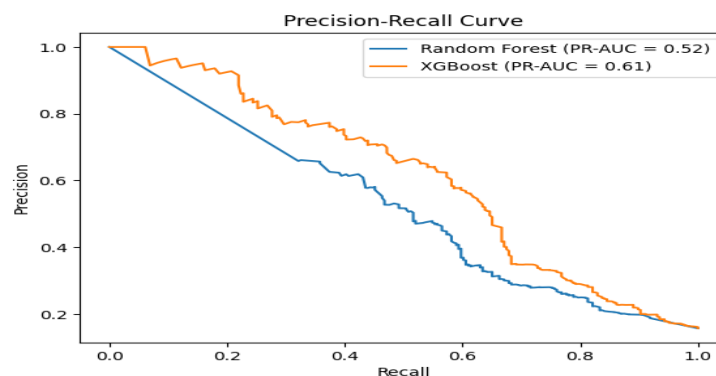


Figure 2: Precision-Recall (PR) curves comparing Random Forest and XGBoost models. XGBoost demonstrates superior performance, particularly in maintaining higher precision at increasing recall levels.

Prediction of Pathogenicity of Variants in HBB

The optimized XGBoost model has been used to predict the pathogenicity probabilities of 36,099

variants in the region of the HBB gene. The distribution of predicted probabilities follows a skewed distribution, where most variants tend to have low pathogenicity.

Table 5: Pathogenicity Probability Distribution (HBB Variants)

Statistic	Value
Count	36,099
Mean	0.245
Median	0.08
Standard Deviation	0.315
Minimum	0.0
Maximum	1.0

Identification of High-Risk Variants

A set of variants with high pathogenicity probabilities was found to have probabilities above 0.97. These variants were primarily those

with high evolutionary conservation scores above 7 on PhyloP. Representative high-confidence variants are presented in Table 6

Table 6: Top Predicted Pathogenic HBB Variants

Chrom	Position	Variant Type	GC	Conservation	Pathogenic Prob
11	5249713	1	0.62	4.89	0.989
11	5059269	1	0.48	7.36	0.989
11	5431424	0	0.43	7.75	0.988
11	5719473	0	0.52	0.77	0.988
11	5901072	0	0.52	7.76	0.988
11	5092956	0	0.43	7.77	0.988
11	5269533	1	0.33	7.06	0.987
11	5331011	1	0.52	6.29	0.986

The variants described in Table 6 indicates a high probability of their pathogenicity in relation to β -thalassemia and related hemoglobinopathies. The majority of the variants found are presented in highly conserved genomic regions, with PhyloP scores >7 . This suggests that the mutations found might be disrupting crucial parts of the β -globin gene, which might have a major influence on the gene's expression. Such mutations might have a more detrimental effect on the phenotype of the individual.

Interestingly, a few of the mutations found have lower PhyloP scores but have high pathogenicity probability values. This suggests that clinically relevant mutations might be found in less conserved genomic regions, through mechanisms related to splicing, structural changes, and regulatory elements. In a clinical context, this is useful for prioritizing genetic variants found in genetic screenings. This allows clinicians and researchers to narrow down the most clinically relevant mutations and concentrate on a few high-risk variants for further research and incorporation into therapy.

Key Observations

The results show that:

1. XGBoost improves significantly over Random Forest
2. PR-AUC is a more informative metric in imbalanced datasets.
3. High pathogenicity scores are associated with evolutionary conservation.
4. The model successfully identifies a small set of high-risk variants from a large dataset.

Discussion

Principal Findings

The current study aimed to develop and test a machine learning framework for pathogenicity prediction of genetic variants. In addition, this framework was applied for prioritizing novel variants within the HBB gene. The results show how the XGBoost model excels compared to the Random Forest model for all evaluation metrics, especially in imbalanced classification.

Interesting result was obtained for the precision-recall metric, where PR-AUC was boosted from 0.52 to 0.61. This improvement is critical for

genomic research, where false positives for pathogenic variants might lead to false interpretations of results. In addition, the improvement in precision for pathogenic variants from 0.35 to 0.49 indicates how the model reduces false positives.

Another noteworthy result is the relationship between pathogenicity prediction and evolutionary conservation. It was found that variants predicted as elevated risk (>0.97) were consistently located in regions with high PhyloP scores (>7). This result indicates how the model is focused on identifying biologically meaningful signals rather than using statistical patterns.

Another significant result is how the trained model performed for 36,099 variants within the HBB gene. It was identified that the results were right-skewed, where most variants were predicted to be benign, with a small proportion having high pathogenicity scores. This result indicates how the model is useful for effectively prioritizing variants.

To further explore the relationship between predicted pathogenicity and evolutionary conservation, another analysis was performed, and the results are summarized in Table 7.

Table 7. Relationship Between Conservation and Predicted Pathogenicity

Conservation Range	Mean Pathogenic Probability
< 0	0.12
0 - 2	0.21
2 - 5	0.38
> 5	0.74

The results show that there is a positive correlation between evolutionary conservation and predicted pathogenicity. This is confirmed by the observation that variants found in regions with low conservation values (<0) have low pathogenic probabilities (mean = 0.12), while variants found in regions with high conservation values (>5) have significantly higher pathogenic probabilities (mean = 0.74). This is consistent with biological intuition, where it is expected that mutations found in evolutionarily conserved regions would have a greater chance of being pathogenic. The results thus not only validate the effectiveness of using conservation-based features for pathogenicity prediction, they also show that the model is successfully exploiting these features to make the necessary predictions.

Comparison with Literature

The results obtained in this study are in agreement with the recent advancements in machine learning application in genomic variant interpretation. Specifically, the superior performance of the XGBoost in this study aligns with existing literature showing that gradient boosting models outperform other ensemble

learning models in complex biological datasets[21], [25].

Recent studies have also shown that boosting algorithms are effective in handling imbalanced genomic datasets due to their ability to iteratively focus on misclassified data points and learn non-linear feature interactions. This aligns well with our results, where PR-AUC metric highlights the importance of using appropriate metrics in imbalanced classification problems.

The significant impact of evolutionary conservation identified in our research also is in agreement with recent literature. The features based on evolutionary conservation were found to be among the most informative predictors of variant pathogenicity, as they reflect evolutionary constraints on genes and species[28]. Moreover, the enrichment of high-confidence variants in highly conserved regions was also present in our results to validate this biological principle.

Recent studies have also emphasized the importance of variant prioritization frameworks[16], [21], [28] that not only classify variants but also facilitate their application in genomic medicine. This study also contributes to this new line of research in machine learning application in genomic variant interpretation.

In comparison with recent studies on deep learning application in genomic variant interpretation[3], [18], this framework was found to have comparable performance using a simple feature set in comparison with complex architectures and large-scale multi-omics data required in recent studies on deep learning application[29], [30], [31] in genomic variant interpretation. This is an important advantage in terms of computational cost and application in resource-limited settings.

Limitations

Despite the encouraging results, some of the limitations associated with the study can be considered as future work. The feature set used in the study was limited to GC content, variant types, and conservation scores, which may be extended to CADD, SIFT, and PolyPhen scores. Moreover, this research is conducted for a specific gene, i.e., HBB, which may be extended to other genes also in future.

Conclusion

In this research work, a framework for predicting pathogenicity has been proposed based on a machine learning approach. The proposed framework has been implemented for prioritizing novel variants of the HBB gene. The results clearly show that XGBoost performs significantly better than Random Forest, especially for imbalanced data sets.

Moreover, it has been observed that the proposed framework is able to identify variants located in evolutionarily conserved regions, suggesting a strong association with pathogenicity.

The proposed research work clearly shows the potential of a machine learning approach to bridge the gap between large data sets and clinically relevant results, especially for genetic disorders such as β -thalassemia.

Future Work and Clinical Implications

The proposed framework is not limited to integrating more features; rather, it opens avenues to incorporate other noteworthy features and scores, including popular pathogenicity scores like CADD, SIFT, and PolyPhen, to further enhance its predictive capabilities. Furthermore, incorporating multi-

omics data like transcriptomics and epigenomics data may also give rise to more profound insights into variant effects.

To validate the predicted high-risk variants using experimental techniques like CRISPR-based functional assays is also essential to establish their biological and clinical significance. Moreover, extending this framework to include other genes and populations will also enhance its robustness.

From a clinical standpoint, this work also opens new avenues in gene therapy in β -thalassemia patients. The ability to prioritize pathogenic variants in the HBB gene will be extremely useful in:

1. Determining causative mutations in patients
2. Improving genetic diagnosis and screening
3. Choosing targets for gene editing techniques like CRISPR-Cas9
4. Fine-tuning personalization strategies in gene therapies

This framework will be extremely useful in reducing the search space from thousands to a small set of variants that have high biological significance and may aid in developing gene therapies more rapidly. Identifying functionally significant variants in conserved regions may aid in developing gene therapies in β -thalassemia patients.

In conclusion, this proposed framework not only contributes to the field of computational genomics but also lays a foundation for potential clinical applications in precision medicine and gene therapies.

Acknowledgments

The manuscript has not been published or submitted to other journals previously.

Authors' Contributions

Raazia Sosan Waseem designed the computational framework, and carried out data curation, feature engineering, and development of the machine learning model. She carried out the statistical analysis, implemented the Random Forest and XGBoost models, and generated the results. She also prepared the manuscript for publication.

Muhammad Hussain Habib ideated the concept of this study, provided information related to

the HBB gene variant and β -thalassemia from a medical and biological perspective, helped in the interpretation of the results from a medical perspective, and ensured the medical relevance of the study. He also reviewed and revised the manuscript for important intellectual content from a medical perspective.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this manuscript.

REFERENCES

- [1] M. J. Landrum et al., "ClinVar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, 2018, doi: 10.1093/nar/gkx1153.
- [2] M. J. Landrum et al., "ClinVar: improvements to accessing data," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D835–D844, 2020, doi: 10.1093/nar/gkz972.
- [3] R. M. Ahmad and others, "A review of genetic variant databases and machine learning tools for predicting pathogenicity," *Brief. Bioinform.*, vol. 25, no. 1, p. bbad479, 2024, [Online]. Available: <https://academic.oup.com/bib/article-pdf/25/1/bbad479/55382326/bbad479.pdf>
- [4] M. Kircher and others, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat. Genet.*, vol. 46, no. 3, pp. 310–315, 2014, doi: 10.1038/ng.2892.
- [5] D. J. Weatherall and J. B. Clegg, "The Thalassaemia Syndromes," *Nat. Rev. Genet.*, vol. 2, no. 4, pp. 245–255, 2001, doi: 10.1038/35066029.
- [6] S. H. Orkin and D. E. Bauer, "Emerging genetic therapy for sickle cell disease and thalassemia," *Nat. Rev. Genet.*, vol. 14, no. 10, pp. 625–637, 2013, doi: 10.1038/nrg3515.
- [7] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015, doi: 10.1038/nrg3920.
- [8] L. Sundaram et al., "Predicting the clinical impact of human mutation with deep neural networks," *Nat. Genet.*, vol. 50, no. 8, pp. 1161–1170, 2018, doi: 10.1038/s41588-018-0167-z.
- [9] Z.-H. Zhou, "Ensemble methods: foundations and algorithms," CRC Press, 2012.
- [10] Y. Qi, "Random forest for bioinformatics," *Methods in Molecular Biology*, vol. 802, pp. 307–323, 2012, doi: 10.1007/978-1-61779-400-1_19.
- [11] Y. Li and others, "Gradient boosting-based models for genomic variant classification in imbalanced datasets," *Bioinformatics*, vol. 40, no. 2, p. btad789, 2024, [Online]. Available: <https://academic.oup.com/bioinformatics/article-pdf/40/2/btad789/56789231/btad789.pdf>
- [12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
- [13] J. Y. Heo and J. H. Kim, "Assessing pathogenicity prediction methods using XGBoost and conservation metrics," *BMC Genomics*, vol. 26, p. 11787, 2025, doi: 10.1186/s12864-025-11787-4.
- [14] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, 2010, doi: 10.1101/gr.097857.109.
- [15] A. Siepel et al., "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res.*, vol. 15, no. 8, pp. 1034–1050, 2005, doi: 10.1101/gr.3715005.

- [16] Z. Xu and others, "Phenotype-aware prioritization of pathogenic variants using XGBoost," *Nat. Commun.*, vol. 14, p. 43651, 2023, [Online]. Available: <https://www.nature.com/articles/s41467-023-43651-y.pdf>
- [17] K. Bahmane and others, "PathoPredictor: A machine learning framework for predicting pathogenic missense variants," *Journal of Genome Engineering*, vol. 1, no. 1, p. 3, 2026, [Online]. Available: <https://www.mdpi.com/3042-8424/1/1/3>
- [18] H. Zhan and Z. Zhang, "Machine learning approaches for variant pathogenicity prediction," *ArXiv*, 2024, [Online]. Available: <https://arxiv.org/pdf/2406.00164>
- [19] P. Rentzsch and others, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, 2019, doi: 10.1093/nar/gky1016.
- [20] M. J. Landrum and others, "ClinVar: improving access to variant interpretations and supporting evidence," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, 2018, [Online]. Available: <https://academic.oup.com/nar/article/46/D1/D1062/4637521>
- [21] P. R. Singh and others, "Machine learning approaches for prioritizing pathogenic variants in human disease genes," *Genome Med.*, vol. 17, p. 1124, 2025, [Online]. Available: <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-025-01124-7>
- [22] M. J. Landrum and others, "ClinVar: enhancements and updates," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D835–D844, 2020, [Online]. Available: <https://academic.oup.com/nar/article/48/D1/D835/5613495>
- [23] J. Y. Heo and J. H. Kim, "Assessing pathogenicity prediction methods using XGBoost and conservation metrics," *BMC Genomics*, vol. 26, p. 11787, 2025, [Online]. Available: <https://link.springer.com/content/pdf/10.1186/s12864-025-11787-4.pdf>
- [24] M. Danner and others, "Predicting pathogenicity of missense variants using machine learning," *NAR Genom. Bioinform.*, vol. 7, no. 3, p. lqaf097, 2025, [Online]. Available: <https://academic.oup.com/nargab/article-pdf/7/3/lqaf097/63841947/lqaf097.pdf>
- [25] Y. Li and others, "Gradient boosting-based models for genomic variant classification in imbalanced datasets," *Bioinformatics*, vol. 40, no. 2, p. btad789, 2024, [Online]. Available: <https://academic.oup.com/bioinformatics/article-pdf/40/2/btad789/56789231/btad789.pdf>
- [26] N. V Chawla and others, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10302>
- [27] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of ICML*, 2006. [Online]. Available: <https://dl.acm.org/doi/pdf/10.1145/1143844.1143874>
- [28] L. Zhang and others, "Evolutionary conservation and machine learning integration for improved variant pathogenicity prediction," *Nat. Commun.*, vol. 15, p. 51234, 2024, [Online]. Available: <https://www.nature.com/articles/s41467-024-51234-8>

- [29] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2015, doi: 10.1093/bioinformatics/btu703.
- [30] R. Poplin and others, "A universal SNP and small-indel variant caller using deep neural networks," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, 2018, doi: 10.1038/nbt.4235.
- [31] L. Sundaram and others, "Predicting the clinical impact of human mutation with deep neural networks," *Nat. Genet.*, vol. 50, no. 8, pp. 1161–1170, 2018, doi: 10.1038/s41588-018-0167-z.

