BIOINFORMATICS

AI-Based Genetic Disease Prediction Using Machine Learning

Pratikshya Shivhari Shinde

Navsahyadri Group of Institutions, Naigaon, Pune, Maharashtra, India

Published November 2026

Abstract

Genetic disorders are caused by changes in DNA that may alter gene function and contribute to disease. Interpreting the large number of variants identified through genetic testing is challenging because variant effects depend on several factors, including sequence context, population frequency, conservation and existing clinical evidence. This paper presents a machine-learning framework for prioritising potentially disease-associated genetic variants. The proposed pipeline combines variant-level information with gene and clinical annotation, applies data cleaning and feature encoding, and compares baseline classification algorithms such as logistic regression, random forest and gradient-boosted decision trees. Model performance is to be assessed using sensitivity, specificity, precision, F1-score, receiver operating characteristic area under the curve (ROC-AUC), precision–recall area under the curve (PR-AUC), and calibration. To reduce overly optimistic estimates, the framework recommends gene-aware or variant-aware data splitting, class-imbalance handling within the training folds, and independent validation. Predictions are intended to support research prioritisation rather than replace clinical interpretation. The system will require validation against curated variant classifications and expert review before any clinical use. This work provides a reproducible design for investigating how machine learning can assist genetic variant prioritisation while addressing data leakage, bias, interpretability and ethical concerns.

Keywords: Genetic disease prediction; machine learning; genomic variants; bioinformatics; variant prioritisation; explainable AI

1 Introduction

Genetic diseases arise from changes in the genome that can affect protein structure, gene expression or biological pathways. Advances in DNA sequencing have increased the number of genetic variants available for analysis.

However, identifying which variants are relevant to a disease remains difficult. Many observed variants are benign or have uncertain significance, and the interpretation of a variant requires evidence from multiple sources.

Machine learning (ML) can identify patterns in labelled data and combine heterogeneous features to estimate whether a genetic variant is likely to be associated with disease. Potentially useful inputs include allele frequency, evolutionary conservation, predicted protein impact, variant consequence, gene-level information and previously curated clinical classifications. Such predictions can help researchers prioritise candidates for further investigation.

The American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) provide standards for interpreting sequence variants using evidence categories and expert assessment [1].

Computational predictions are one evidence type and should not be treated as a standalone clinical diagnosis. Recent work also emphasises the importance of calibration, robustness and transparent assessment when machine-learning variant-effect predictions are considered for clinical interpretation [2,3].

This paper proposes a structured ML pipeline for genetic disease variant prioritisation. Its main contributions are:

(i) a modular data-processing and feature-engineering workflow; (ii) comparison of interpretable and ensemble ML models; (iii) an evaluation protocol designed to reduce data leakage and class-imbalance effects; and (iv) safeguards for responsible use and interpretation.

2 Problem statement and objectives

The problem addressed is the prioritisation of genetic variants as potentially disease-associated or likely benign using curated variant annotations. The task is framed as a binary classification problem for an initial prototype. In real-world genetics, clinical classifications may include five categories (benign, likely benign, uncertain significance, likely pathogenic and pathogenic); collapsing these categories into two classes must be justified and documented.

The objectives are to identify informative variant features, train and compare suitable ML classifiers, evaluate their predictive performance on data not used for training, and provide interpretable outputs that communicate uncertainty. The system is a research decision-support prototype and is not intended to diagnose individuals.

3 Literature review

Richards et al. established ACMG/AMP recommendations for sequence-variant interpretation and described how computational evidence can contribute to a broader evidence framework [1]. These recommendations make clear that in-silico predictions are only one component of variant interpretation and must be considered with other evidence.

Duong and Solomon reviewed applications of artificial intelligence in clinical genetics, including diagnostic support and clinical decision-making, while identifying important opportunities and limitations [2]. Dias et al. discussed approaches for evaluating and calibrating computational predictions of variant effects so that predictions can be used more reliably as evidence [3].

Traditional ML models, including logistic regression, support vector machines, random forests and gradient boosting, offer different trade-offs between interpretability, flexibility and predictive performance. Deep-learning methods can learn complex representations from sequence or protein data, but often require larger datasets, careful validation and more substantial computational resources. For a student-level prototype, classical ML models provide a practical and reproducible starting point.

A recurring challenge in genomic prediction is dataset bias. Variants from the same gene, family or highly similar genomic regions may appear in both training and test sets, allowing a model to learn

dataset-specific patterns rather than generalisable biological relationships. Evaluation should therefore include leakage-aware splitting and, where possible, independent external validation.

4 Proposed methodology

4.1 Overall system architecture

The proposed system consists of six stages: (1) data acquisition from a documented public, curated source; (2) data cleaning and quality control; (3) feature extraction and encoding; (4) model training and hyperparameter tuning; (5) evaluation on held-out data; and (6) interpretation and report generation.

The resulting workflow is: curated variant data → quality control → feature engineering → training/validation/test split → ML model training → performance evaluation → interpretable prioritisation output.

4.2 Dataset selection

The study should use a public, versioned and appropriately licensed dataset containing genetic variants and expert-curated clinical significance labels. ClinVar is a potential source of submitted interpretations of human genetic variants [4]. The dataset release date, download date, access method, filtering criteria and label definitions must be recorded. The authors must check the dataset's terms of use and document the evidence review status of included records.

For an initial binary experiment, pathogenic and likely pathogenic records may form the positive class, while benign and likely benign records may form the negative class. Variants classified as uncertain significance, conflicting interpretations or having insufficient evidence should be excluded from the primary binary experiment or analysed separately. The number of records in each category must be reported after filtering; it must not be estimated or invented.

4.3 Data preprocessing

Duplicate records should be identified using a consistent variant identifier and genome assembly. Records with missing essential labels or invalid variant fields should be removed according to predefined rules. Missing feature values should be handled using imputation fitted only on the training data. Categorical variables should be encoded using a documented method, and numerical variables should be scaled when required by the selected algorithm. All preprocessing operations must be fitted within each training fold to prevent information leakage.

4.4 Feature selection and extraction

Candidate features may include: (i) variant consequence (for example, missense, synonymous or stop-gained); (ii) allele frequency in reference populations; (iii) evolutionary conservation scores; (iv) predicted protein-impact scores from established tools; (v) gene-level constraint or functional annotations; and (vi) available clinical evidence fields. Only features available at prediction time should be used. The target label itself, post-classification fields, review-status fields that directly reveal the label, and duplicate representations of the same annotation must not be used as predictors.

Feature selection should be performed only within training data. Correlation analysis, permutation importance or model-based feature importance can be used, but feature importance should be interpreted as model behaviour rather than proof of biological causation.

4.5 Machine-learning models

Three baseline models are proposed. Logistic regression provides a relatively interpretable linear baseline. Random forest can model nonlinear relationships and interactions among features. Gradient-boosted decision trees can provide a strong tabular-data baseline. A support vector machine may be added if the dataset size and feature representation make it practical. Model choice must be justified by the available data and computing resources.

Hyperparameters should be selected using cross-validation on the training data. If classes are imbalanced, class weights or resampling methods may be tested within training folds only. The final test set must remain untouched until model selection is complete.

4.6 Training and validation strategy

The dataset should be divided into training, validation and test subsets, for example 70%, 15% and 15%, provided that the sample size supports this division. A random stratified split may be reported as a baseline, but a stronger evaluation should group related records by gene or another biologically justified unit so that closely related variants do not occur across both training and test sets. An independent external dataset is recommended for a further generalisation test.

The random seed, software versions, feature definitions, inclusion criteria and split strategy should be reported to support reproducibility. If multiple records correspond to the same variant or submitter evidence, the authors should prevent those records from leaking across data partitions.

4.7 Evaluation metrics

Performance should be assessed using the confusion matrix and the following metrics: accuracy, precision, sensitivity (recall), specificity, F1-score, ROC-AUC and PR-AUC. PR-AUC is particularly informative when the positive and negative classes are imbalanced. Probability calibration should also be examined if the system outputs risk scores. Metrics should be reported with confidence intervals where feasible, and results should be compared at a clearly stated decision threshold.

The metrics are defined in Equations (1)–(5):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively.

5 Results

This section must be completed after the proposed pipeline has been implemented and evaluated. No experimental measurements are asserted in this draft. Insert the actual dataset counts, model

outputs and plots generated from the experiment. Do not submit blank tables or replace missing results with assumed values.

Table 1 presents the test-set performance comparison.

Table 1. Test-set performance comparison.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	[Add]	[Add]	[Add]	[Add]	[Add]
Random Forest	[Add]	[Add]	[Add]	[Add]	[Add]
Gradient Boosting	[Add]	[Add]	[Add]	[Add]	[Add]

Note: Replace every [Add] field with a measured value and state the test-set size and confidence intervals in the accompanying text.

The Results section should also report class distribution, confusion matrices, ROC and precision–recall curves, calibration results (if probabilities are used), and the outcome of any external validation. Include a short error analysis of false-positive and false-negative predictions.

6 Discussion

The comparison should examine whether more complex models provide meaningful improvements over the interpretable baseline, and whether those improvements remain under gene-aware or external validation. A high score on a random split alone is insufficient evidence of clinical usefulness. The discussion should consider class imbalance, uncertain labels, dataset ascertainment bias, missing annotations, population representation and differences in evidence quality.

Feature-importance methods can help explain which annotations influence predictions, but they do not establish causal mechanisms. A model may reproduce biases in its training labels or perform differently across ancestry groups, genes or disease categories. Subgroup performance should be assessed where sample sizes and ethical approvals permit. Any apparent performance differences should be interpreted cautiously.

The proposed framework is intended to prioritise variants for further research. It does not establish a diagnosis, determine an individual's future health, or replace genetic counselling, laboratory confirmation or expert clinical interpretation.

7 Limitations

The proposed framework depends on the completeness and quality of curated variant annotations. Clinical labels can change as new evidence becomes available, and some variants remain uncertain. Feature availability may vary by gene, population and annotation pipeline. A binary label scheme simplifies the richer clinical classification system and can conceal uncertainty. Model performance may also be affected by class imbalance, correlated records and dataset shift. These limitations must be discussed using the actual dataset and experiment rather than assumed away.

8 Ethical considerations and responsible use

Genetic information is sensitive and can have implications for individuals and biological relatives. This prototype should use public or appropriately authorised data and should not process identifi-

able patient data without institutional approval, consent where required, and suitable data protection. The authors should document dataset licences and any ethics review or exemption. Predictions must be communicated with uncertainty and must not be presented as medical advice. Clinical deployment would require independent validation, expert oversight, privacy safeguards and applicable regulatory review.

9 Conclusion

This paper presents a reproducible machine-learning framework for prioritising potentially disease-associated genetic variants. The proposed workflow combines curated variant annotations, careful preprocessing, baseline classification models and leakage-aware evaluation. It also highlights the need for calibration, interpretability, subgroup assessment and expert review. The next step is to implement the pipeline on a documented dataset, report measured results and validate the selected model on independent data. Until such validation is completed, the framework should be regarded as a research prototype rather than a clinical diagnostic system.

10 Future scope

Future work may incorporate DNA and protein sequence representations, gene–disease knowledge graphs, multimodal clinical evidence and calibrated ensemble models. Further studies should evaluate performance across ancestries, disease groups and independent clinical cohorts. Prospective assessment with genetic specialists would be necessary before considering clinical integration.

References

[1] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–424. doi:10.1038/gim.2015.30

[2] Duong D, Solomon BD. Artificial intelligence in clinical genetics. *Eur J Hum Genet*. 2025;33:281–288. doi:10.1038/s41431-024-01782-w

[3] Dias M, Orenbuch R, Marks DS, Frazer J. Toward trustable use of machine learning models of variant effects in the clinic. *Am J Hum Genet*. 2024;111(12):2589–2593. doi:10.1016/j.ajhg.2024.10.011

[4] National Center for Biotechnology Information. *ClinVar: A public archive of interpretations of clinically relevant variants* [Internet]. Bethesda (MD): NCBI. Available from: https://www.ncbi.nlm.nih.gov/clinvar/ [access date to be supplied by the author]

[5] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017); 2017. p. 4765–4774.