Predictive Modelling for Genetic Data Significance: An Experimental Study

Bhushan Yelure, Suraj Kadam, Vinayak Thorat, Rohan Mali, Sourabh Patil and Neha Pawar

ABSTRACT

Identification of genetic variations associated with complicated characteristics, including susceptibility to particular diseases, has become a crucial function of Genome-Wide Association Studies (GWAS). However, the challenge of predicting statistical Single Nucleotide Polymorphisms (SNPs) remains due to high-dimensional data and complex genetic architectures. In this study, seven regression models are assessed: Random Forest, Extra Trees Regressor, XGBoost, CatBoost, LightGBM, Support Vector Regressor, and Elastic Net for their predictive capacity and precision in predicting SNP significance using transformed *P*-values (–log10 scale) as a continuous regression target, with Bonferroni and False Discovery Rate corrections applied during preprocessing to define significance thresholds for interpretation. By transforming *P*-values using a logarithmic scale, we examine model performance based on various metrics, including mean squared error, mean absolute error, explained variance and R-squared, to determine which models best predict SNP significance. Research findings show that tree-based ensemble methods, particularly the Random Forest and Extra Trees Regressor, achieve the highest predictive accuracy, with Random Forest emerging as the top performer. Gradient boosting models, such as XGBoost and CatBoost, also demonstrate robust results, indicating their ability to capture complex SNP interactions. The research study provides insight into the selection of the model for GWAS-based predictions and contributes to methodologies for more accurately identifying SNPs with potential implications in disease risk prediction. The results obtained offer practical guidance for researchers in choosing appropriate regression models for high-dimensional genetic data.

Keywords: SNP significance prediction, Ensemble regression GWAS, Gradient boosting trees, Bonferroni-FDR adjustment, Varicose veins genetics

Department of Computer Science & Engineering, Government College of Engineering, Kolhapur, India

Correspondence to:
Bhushan Yelure,
bhushan.yelure@gcekarad.ac.in

Additional material is published online only. To view please visit the journal online.

Cite this as: Yelure B, Kadam S, Thorat V, Mali R, Patil S and Pawar N. Predictive Modelling for Genetic Data Significance: An Experimental Study. Premier Journal of Science 2025;15:100225

DOI: https://doi.org/10.70389/PJS.100225

Peer Review

Received: 14 August 2025

Last revised: 6 October 2025

Accepted: 17 December 2025

Version accepted: 2

Published: 29 January 2026

Ethical approval: N/a

Consent: N/a

Funding: N/a

Conflicts of interest: N/a

Author contribution: Bhushan Yelure, Suraj Kadam, Vinayak Thorat, Rohan Mali, Sourabh Patil and Neha Pawar – Conceptualization, Writing – original draft, review and editing

Guarantor: Bhushan Yelure

Introduction

Genome-Wide Association Studies (GWAS) are essential for determining genetic loci that influence complex traits and disease susceptibilities, helping to elucidate the genetic architecture of human health and disease. These studies often involve analyzing the genome, which contains millions of Single Nucleotide Polymorphisms (SNPs), generating highdimensional data that necessitate robust and scalable predictive models. A common challenge in GWAS is distinguishing statistically significant SNPs from vast datasets, given the stringent multiple testing corrections required to control false positives, often achieved via the Bonferroni adjustment.[1] However, determining true significance amidst massive genomic data remains computationally intensive and methodologically challenging. In recent years, machine learning models have shown promise in predicting outcomes from complex datasets, particularly with the use of ensemble methods and boosting techniques that can capture non-linear relationships within high-dimensional data.[2] Ensemble learning models such as Random Forest and Extra Trees Regressor are well regarded for their interpretability and ability to handle high dimensional, heterogeneous data, while boosting algorithms like XGBoost and LightGBM (LGBM) have proven effective in capturing subtle interactions between variables through iterative refinement.[3,4] Additionally, Support Vector Machines (SVMs) and linear models, including Elastic Net (EN), have been applied in GWAS, offering valuable comparisons in model performance across diverse algorithmic approaches.[5,6] The ability of seven machine learning models to regress transformed *P*-values (–log10) is carefully compared, providing a quantitative basis for evaluating how well each model captures variation in SNP significance. To stabilize predictive performance, *P*-values will be converted to a logarithmic scale. Multiple evaluation metrics are applied, including Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), and Explained Variance, to identify the most accurate and robust models for SNP significance prediction. It provides a comprehensive assessment of each model's strengths and limitations in genetic data prediction and offer insights for researchers on selecting suitable predictive models for high dimensional genomic studies. The technique employed, comparative model performance results, and an attempt to analyze the findings' significance for the field of genetic epidemiology and beyond are covered in length in the parts that follow.

Related Work

Machine learning's use in GWAS is a burgeoning field, with several studies demonstrating the utility of various models for SNP significance prediction. Traditional statistical approaches in GWAS, such as linear regression, often struggle with the complexity and dimensionality of genomic data, necessitating innovative machine learning methodologies to enhance predictive accuracy and interpretability.[7]

Machine Learning in GWAS and Genomic Prediction

Numerous research have emphasized ensemble learning methods like Random Forest and Extra Trees Regressor, which are well-liked because of their capacity to grasp nonlinear correlations and their resilience when working with big datasets.[3] For example,

Provenance and peer-review: Unsolicited and externally peer-reviewed

Data availability statement: N/a

Chen and Ishwaran[8] demonstrated that Random Forest could significantly improve the accuracy of genetic trait prediction by combining multiple decision trees, thus accounting for interactions across SNPs and achieving higher interpretability in model outputs.[8] Boosting algorithms, particularly XGBoost and LGBM, have emerged as cutting edge approaches in various predictive tasks, including genetic studies, where they are well-suited for dealing with class imbalance and sparse data structures.[2,4] Because of its regularization processes and efficient bias and variance minimization, XGBoost has been demonstrated to perform better in comparative experiments than other machine learning models and conventional statistical methodologies. Likewise, CatBoost, an algorithm specifically optimized for categorical features, has been employed in genetic prediction tasks, further illustrating the adaptability and performance of boosting algorithms in high-dimensional genomic datasets.[9]

Linear and Non-Linear Model Comparison in Genomic Data

While ensemble models dominate the field, linear models like EN remain valuable, particularly for interpretability and feature selection in genomic studies. EN's effectiveness stems from its combination of Lasso and Ridge regression penalties in managing multicollinearity among SNPs and selecting sparse feature sets.[5] Studies have shown that EN can be a complementary tool in GWAS, especially when used alongside more complex models to identify a subset of highly predictive SNPs.[10] SVMs, though less commonly applied in highdimensional genomic data due to scalability concerns, have also been shown to perform well in specific applications, especially in binary classification tasks related to genetic marker identification.[6] SVMs offer an advantage in datasets where the number of features far exceeds the number of samples, a frequent scenario in GWAS, by maximizing the margin of separation between classes.

Comparative Evaluations of Machine Learning Models

Recent comparative studies have provided insights into the relative performance of various machine learning models in GWAS. For example, a study by Goldstein et al.[11] evaluated the predictive accuracy of Random Forest, EN, and SVM for trait prediction, demonstrating that ensemble models like Random Forest consistently outperformed linear models and SVM in handling the non-linear dependencies within genomic data.[11] Similarly, Wright et al.[12] compared the performance of gradient boosting methods and linear models, finding that boosting algorithms achieved higher accuracy across multiple evaluation metrics, particularly in large-scale datasets.[12] Current research study builds upon existing research by conducting an extensive evaluation of seven regression models—Random Forest, Extra Trees Regressor, XGBoost, CatBoost, LGBM, Support Vector Regressor (SVR), and EN using metrics such as MSE, R-squared, and MAE. Research work contributes to the field by identifying model strengths and limitations, offering researchers practical guidance in selecting suitable algorithms for high-dimensional SNP significance prediction.

Methodology

Data Collection

The study utilizes a high-quality GWAS dataset containing SNPs and associated *P*-values, collected from publicly accessible dataset by University of Oxford. The dataset, representing genetic variations across multiple chromosomes, provides the necessary information on SNP positions, allele frequency (A1FREQ), effect sizes (BETA), standard errors (SE), and *P*-values (PVAL). These variables are critical for assessing each SNP's impact and significance concerning specific genetic traits. High-quality SNPs are filtered based on criteria such as information score (INFO > 0.9), a commonly used metric for assessing imputation quality and reliability in genetic studies.[13]

Data Exploration

The initial step of data exploration involved statistical analysis of SNP distributions, allele frequencies, and BETA to characterize the dataset's structure and variability. The *P*-value distribution is visualized through Manhattan and Quantile–Quantile (QQ) plots to assess significant SNPs' dispersion and identify deviations from expected distributions (Figure 1). Manhattan

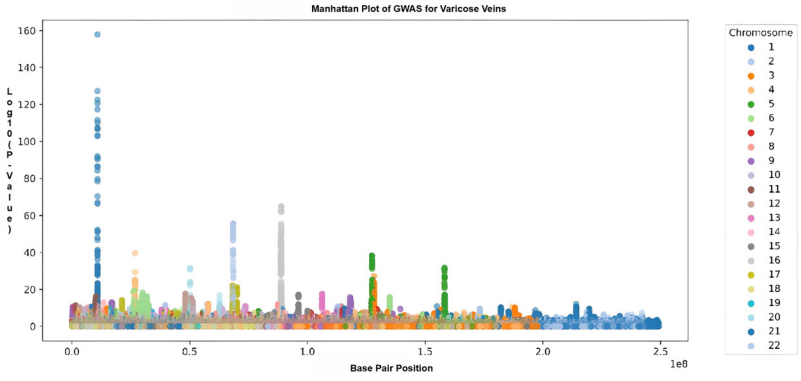

Fig 1 | Manhattan plot of GWAS for varicose veins, illustrating genome-wide SNP significance

DOI: https://doi.org/10.70389/PJS.100225 | Premier Journal of Science 2025;15:100225

plot providing a genome-wide overview of SNP significance. Peaks represent genomic regions with potential associations (Figure 2). QQ plot assessing systematic inflation or deflation in *P*-values, indicating deviations from expected null distributions.[14]

Data Preprocessing

Transformation of* P-*Values:

Given that *P*-values are often skewed in GWAS data, a transformation is applied to make the data more conducive to regression analysis. *P*-values are converted to their negative log base 10 (log_10(*P*)) to stabilize the variance, allowing the model to handle significant and non-significant SNPs on a comparable scale.

Bonferroni Correction and Labelling:

To identify statistically significant SNPs, Bonferroni correction is applied, adjusting the significance level determined by the total number of tests to control for type I errors. SNPs below the Bonferroni-adjusted threshold are labelled as "case" (significant), and those above as "control" (non-significant). This binary labelling is applied during preprocessing to categorize SNPs for interpretative and thresholding purposes. However, the predictive task itself remains regression, using the transformed *P*-values (–log10) as the continuous target variable.

False Discovery Rate (FDR) Using Benjamini-Hochberg (BH) Procedure:

The BH procedure adjusts *P*-values to control the FDR, providing a less conservative alternative to Bonferroni correction while maintaining statistical power. SNPs with FDR adjusted *P*-values below a set threshold (e.g., 0.05) are labelled as "case" (statistically significant), indicating potential association with the phenotype, while those above the threshold are labelled as "control" (non-significant). This binary labelling allows for supervised learning models to differentiate between significant and non-significant SNPs. By prioritizing the control of false discoveries over stringent error rates, the BH procedure ensures a balanced approach to SNP selection in GWAS studies.

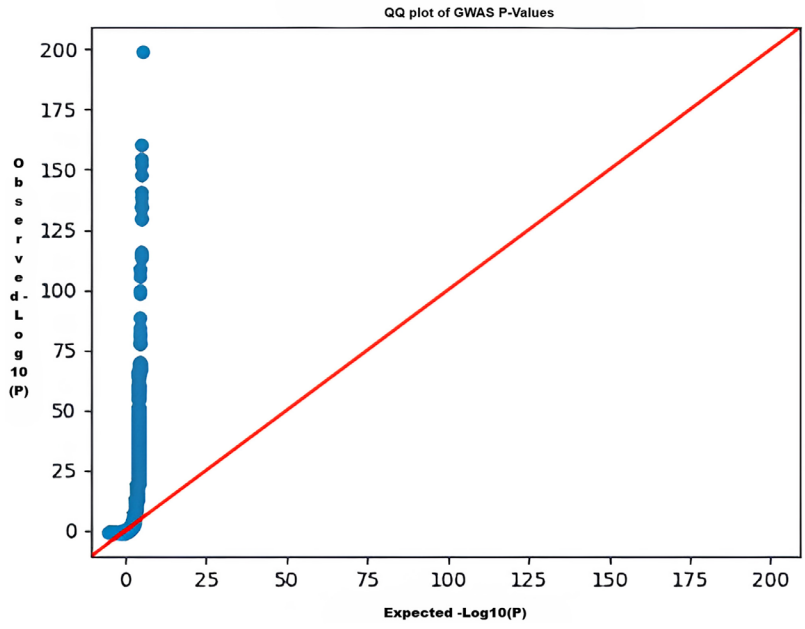

Fig 2 | QQ plot evaluating SNP significance by comparing observed versus expected *P*-values, identifying possible inflation or deflation in test statistics

Feature Selection and Data Splitting:

Based on their predictive power and biological relevance, features such A1FREQ, BETA, SE, and genomic position (chromosome, base pair) are chosen. Training and testing sets are separated from the data, usually in an 80–20 split, over subset comprising of 2 lakh records having highest *P*-value, out of 89,44,547 total available records, to assess model performance accurately. This ensures that most significant genomic data is being used to train the regression models.

Data Sub Setting Strategy and its Implications

For this study, a subset comprising the top 200,000 SNPs with the most significant *P*-values was selected from the original dataset of 8.9 million records. This design choice was driven by computational resource constraints and the objective of conducting a focused, exploratory evaluation of different machine learning models on the most relevant genomic signals. Consequently, both training and testing were performed within this enriched subset, which naturally inflates performance metrics such as R^2 because the variance in this tail of the distribution is lower and the signal-to-noise ratio is higher. While this approach is not intended to reflect genome-wide predictive performance, it provides a consistent evaluation environment for comparing models relative to one another within the same significance spectrum. Future work will extend this evaluation to unbiased or chromosome-wise splits to assess genome-wide generalization.

In the original GWAS framework, the *P*-value is mathematically derived from the BETA and SE using the test statistic t = BETA/SE. Including BETA and SE as predictors while using the transformed *P*-value (–log10 *P*-value) as the target would lead to circularity, allowing the models to reconstruct the *P*-value formula rather than uncover novel genomic patterns. To address this, BETA and SE were excluded from the predictor set. Instead, only independent SNP-level features such as A1FREQ, chromosome number (CHR), base pair position (BP), and imputation information score (INFO) were used as model inputs. This step ensures that the models rely on genomic features rather than the mathematical definition of the target variable, reducing artificial inflation of performance metrics.

Machine Learning Models

Research work covers various machine learning models where, they were employed to evaluate their ability to predict the significance of SNPs based on genomic features. Each model brings unique strengths to

handling regression tasks in genetic data. Here is a brief overview of the models utilized:

Random Forest Regressor (RFR)

To generate a final forecast, an ensemble learning method known as Random Forest builds many decision trees and aggregates their output. It handles non-linear relationships well, is robust to overfitting due to random sampling, and performs well on high-dimensional data. Its feature importance scores also offer insights into which variables contribute most to predictions.[3]

Extra Tree Regressor (EFR)

The algorithm for Extra Trees is an ensemble of decision trees like Random Forest, but differs by splitting on random thresholds rather than best splits. This randomness can improve the model's robustness and generalization ability. ETR is well-suited for handling noise in the data and often exhibits faster training times due to simplified split evaluations.[15]

XGBoost Regressor (XGBR)

XGBoost is a highly optimized gradient boosting method that builds trees one after the other, aiming to correct errors of previous trees. It is effective for large datasets and complex patterns, offering control over regularization, which reduces the risk of overfitting. XGBoost's flexibility makes it popular in genomics, where capturing subtle interactions among SNPs is crucial.[2]

CatBoost Regressor (CBR)

A gradient boosting approach called CatBoost performs well with categorical data and doesn't require a lot of preprocessing. It automatically manages categorical features, has strong performance on diverse datasets, and is less sensitive to hyperparameter tuning compared to other boosting methods. Its boosting framework is suitable for genomic data that may contain mixed data types.[9]

LGBM

LGBM is a gradient boosting framework designed for efficiency and scalability on large datasets. It employs a leaf-wise tree growth strategy with depth constraints, enabling faster training and better accuracy compared to traditional level-wise methods. LGBM is well suited for handling high-dimensional genetic data because it supports sparse input, handles categorical features effectively, and can capture complex non-linear interactions between SNP features. Its ability to train quickly while maintaining strong predictive performance makes it a practical choice for large-scale GWAS regression tasks.

SVR

SVR finds the hyperplane that best fits the data within a tolerance margin, extending SVMs to regression tasks. It is effective for small to medium-sized datasets and can capture complex relationships but may require kernel tuning and is computationally intensive with large data.[16]

EN

EN balances feature selection (L1) and stability (L2) by combining L1 and L2 regularization. It is especially helpful in high-dimensional datasets that have correlated features, as it can handle multicollinearity, making it suitable for SNP data where genomic features are often correlated.[5]

Model Building

The model-building process in this study followed a structured methodology, as shown by Figure 3, designed to rigorously evaluate multiple regression models for predicting significant SNPs in GWAS. This section outlines the main steps in the process, including the key tasks and considerations undertaken for each model:

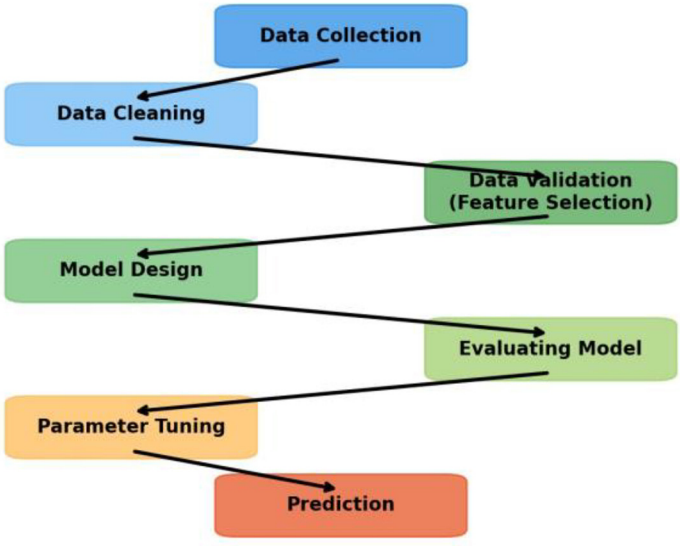

Fig 3 | Flowchart outlining the overall process of the predictive modelling workflow used for SNP significance evaluation

DOI: https://doi.org/10.70389/PJS.100225 | Premier Journal of Science 2025;15:100225

Feature Engineering and Data Preprocessing

- Data Cleaning: Initial data cleaning involved filtering for high-quality SNPs using an INFO score threshold >0.9 to remove lower-quality imputation results. The dataset was then processed to remove duplicates and any SNPs with missing values.
- Feature Selection: Key features influencing SNP significance were selected based on domain knowledge, including A1FREQ, BETA, SE, INFO, CHR, and BP.
- Target Transformation: Since *P*-values often span several magnitudes, they were transformed to (log_10(p)) to improve model stability and interpretability. This transformed (log_pval) became the target variable.
- Class Labelling: SNPs with *P*-values below the Bonferroni threshold were labelled as "significant" (1), while others were labelled as "non-significant" (0), establishing a binary classification target for significance.

Data Splitting and Training/Test Setup

- Train-Test Split: For computational efficiency and reproducibility, the dataset subset was split into training (80%) and testing (20%) sets using a random state of 42.
- Standardization: Continuous features were standardized to ensure consistent model performance, especially for algorithms sensitive to feature scale like SVR and ElasticNet.

Validation Strategy

In this study, model performance was evaluated using a single train–test split (80–20) with a fixed random seed to ensure reproducibility. Hyperparameters were kept at their default settings to maintain a uniform configuration across all models and allow for a fair comparative baseline without introducing variability from extensive tuning. Cross-validation, chromosome hold-out strategies, or permutation testing were not applied in this phase due to computational constraints and the exploratory nature of the analysis. As a result, the reported performance metrics should be interpreted as baseline comparative results rather than fully optimized outcomes. Future work will incorporate more rigorous validation schemes, including k-fold cross-validation, hyperparameter optimization, and chromosome-wise hold-outs to assess generalizability more comprehensively.

Model Selection and Configuration

To offer a varied evaluation of predicted performance, seven distinct machine learning models were selected. Each model was initialized with a default configuration and, where applicable, customized based on the model's unique requirements and the data structure.

- RFR: Utilized for its robustness to overfitting and its effectiveness with highdimensional datasets.
- ETR: Similar configuration as RFR but with randomness in node splitting.
- XGBR: Known for its performance in structured data, it was configured in order to solve the regression problem.
- CBR: Automatically manages categorical variables, useful for mixed feature types.
- LGBM: Leaf-wise tree growth was applied for efficient handling of large data.
- SVR: As SVR is computationally intensive, took a lot of time in order to train the model.
- EN: Leveraging both L1 and L2 regularization to reduce feature multicollinearity.

Model Evaluation

- All models were evaluated using regression-based metrics, as the predictive task involves modeling transformed *P*-values as a continuous variable, not binary classification.
- Metrics: The following evaluation metrics were used to capture different aspects of model performance: MSE, R-squared (R^2), MAE, Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Explained Variance, Median Absolute Error (MedAE), and Mean Squared Log Error (MSLE).
- Rating System: Each metric was rated on a scale of 1–5, following thresholds adapted from prior studies and standard benchmarks for regression performance. For metrics favoring higher values (e.g., R^2, Explained Variance):

  5: $R^2 \geq 0.90$ (Excellent)
  4: $0.75 \leq R^2 < 0.90$ (Good)
  3: $0.50 \leq R^2 < 0.75$ (Moderate)
  2: $0.25 \leq R^2 < 0.50$ (Poor)
  1: $R^2 < 0.25$ (Very Poor)

  For metrics favoring lower values (e.g., MSE, RMSE):
  5: MSE/RMSE ≤ 0.01 (Excellent)
  4: 0.01 < MSE/RMSE ≤ 0.05 (Good)
  3: 0.05 < MSE/RMSE ≤ 0.10 (Moderate)
  2: 0.10 < MSE/RMSE ≤ 0.20 (Poor)
  1: MSE/RMSE > 0.20 (Very Poor)

Model Comparison and Selection

- Performance Ranking: Based on cumulative ratings across all metrics, models were ranked. RFR achieved the highest rating, followed by ETR, XGBR, CBR, LGBM, SVR, and EN respectively.
- Feature Importance Analysis: Feature importance scores from tree-based models (RFR, ETR, XGBR) were analysed to identify influential features for SNP significance.

The rigorous methodology provided a balanced framework to assess each model's effectiveness in predicting SNP significance. Each step, from preprocessing to ranking, was designed to ensure reliable results and draw meaningful conclusions from the models' comparative performance.

Experimental Results

All analyses are conducted within a regression framework, using −log10(*P*-value) as the continuous response variable. Binary significance thresholds (Bonferroni/FDR) are only used for interpretation and labelling during preprocessing, not as predictive outputs. This section details the performance of each model used to predict SNP significance based on regression metrics, comparing accuracy, error, and explainability across different methodologies, as shown in Figure 4.

RFR

With a low MSE of 0.0255 and a high R-squared (R^2) value of 0.9968, the RFR demonstrated exceptional predictive capacity and obtained the top performance scores. This model provided minimal prediction errors and an impressive Explained Variance score of 0.9902, suggesting it effectively captured the variance in SNP significance with minimal overfitting. Consistent predictions are indicated by its low RMSE and MAE. Consequently, RFR ranked highest among models, demonstrating its reliability in SNP prediction tasks.

ETR

With an extremely low MSE of 0.0074 and an R^2 value of 0.9991, the Extra Trees Regressor performed similarly to RFR, demonstrating how accurate its predictions are. The model's RMSE of 0.0862 and MAE of 0.0111 highlight its capacity to produce precise forecasts with few errors. Despite having a little lower MAPE rating of 0.2538, it performed second overall, only surpassed by RFR, making it a robust alternative for SNP significance prediction.

XGBR

The XGBR performed well with an R^2 of 0.9854 and an MSE of 0.1165. With an RMSE of 0.3413 and an MAE of 0.0560, it achieved consistently accurate results.

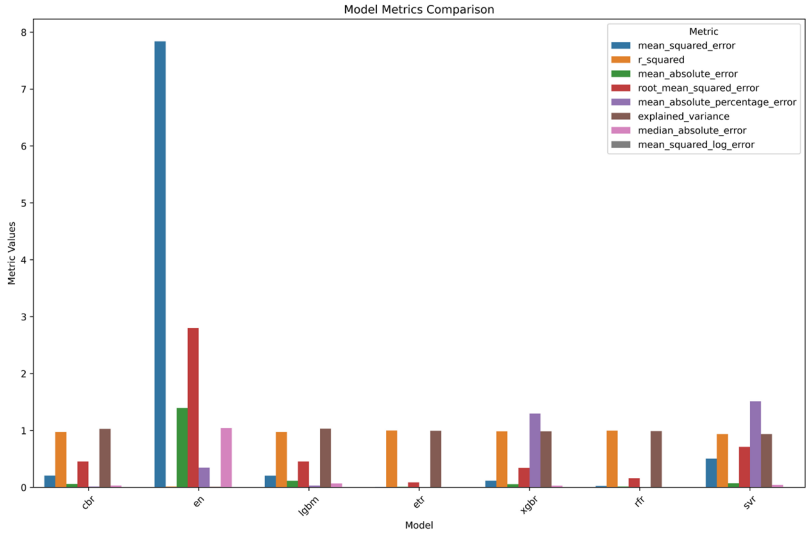

Fig 4 | Comparison of different regression models based on performance metrics such as MSE, MAE, R-squared, and explained variance

XGBR demonstrated strong performance in Explained Variance (0.9854) and MAPE, indicating a balanced performance in predictive accuracy and generalization. Its cumulative metrics place it third overall, reflecting its strength in handling structured genetic data effectively.

CBR

Cat Boost's R^2 score of 0.9741 and MSE of 0.2062 were comparable to those of XGBR. Its RMSE of 0.4540 and MAE of 0.0588 suggest a higher degree of variance in predictions than RFR and ETR. However, with strong scores in Explained Variance (1.0289) and low MAPE, CatBoost remains a solid choice for SNP prediction tasks where categorical features are prevalent.

LGBM

LGBM obtained an MSE of 0.2056 and a respectable R^2 of 0.9741, though its RMSE (0.4534) and MAE (0.1146) were relatively higher, indicating slight prediction variance. Its high Explained Variance score of 1.0314 and low MAPE demonstrate its strength in certain aspects, yet it performed less consistently compared to RFR, ETR, and XGBR.

SVR

With an MSE of 0.5055 and an RMSE of 0.7110, the SVR model obtained an R^2 of 0.9364.Its MAE of 0.0707 indicates satisfactory performance, although its higher MSE and RMSE suggest it might not fully represent the range in SNP data as effectively as other models. SVR, while performing well in some metrics, ranked lower due to its reduced ability to handle the dataset's complexity.

EN

EN exhibited the lowest performance, with a noteworthy high MSE of 7.8385 and an R^2 value of 0.0139. Its high RMSE (2.7997) and MAE (1.3954) indicate substantial error variance. Despite its strength in reducing feature multicollinearity, ElasticNet struggled to capture significant patterns in this dataset (Figure 5).

A subset of 200,000 records was chosen from the entire dataset in order to guarantee computing effectiveness while preserving statistical power. To ensure that the distribution of important variables (such as SNP allele frequencies, phenotypic traits, and demographic factors) in the subset closely reflects that of the entire dataset, the selection process was carried out using a stratified sampling strategy. This method preserves the generalizability of the dataset by avoiding the overrepresentation or underrepresentation of particular subgroups. Because models were trained and evaluated using a fixed train–test split and default hyperparameters, the results primarily reflect comparative baseline performance rather than fully optimized predictive capabilities. These choices enable a consistent evaluation framework across models but may not capture the full potential of each algorithm under optimized settings.

DOI: https://doi.org/10.70389/PJS.100225 | Premier Journal of Science 2025;15:100225

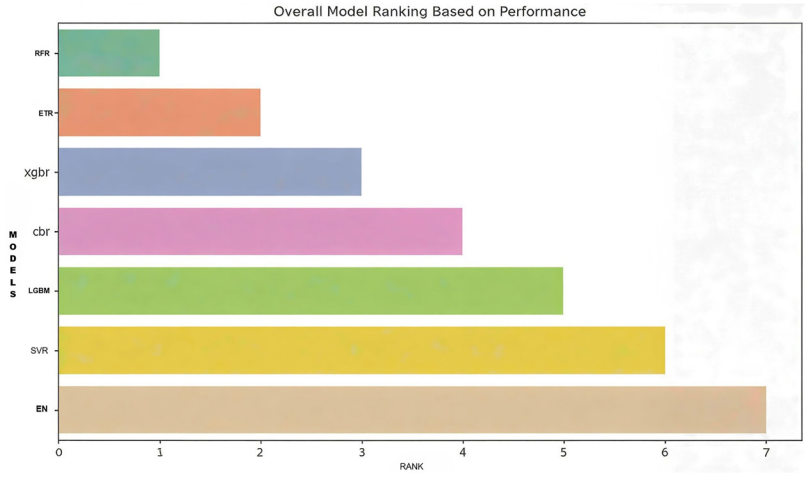

Fig 5 | Performance-based ranking of regression models, demonstrating which models outperformed others based on evaluation criteria like MSE, MAE, R-squared, and explained variance

Table 1 | Performance metrics of ml models based on MSE, MAE, and R^2 values

ML Models	MSE (bp^2)	MAE (bp)	R^2 Value
RFR	0.2550	0.0139	0.9968
ETR	0.0074	0.0111	0.9991
XGBR	0.1165	0.0560	0.9854
CBR	0.2062	0.0588	0.9740
LGBM	0.2056	0.1146	0.9740
SVR	0.5055	0.0707	0.9364
EN	7.8385	1.3954	0.0139

Table 2 | Genetic association statistics for top significant SNPs

SNP	PVAL	BETA	A1FREQ	INFO
rs11121615	1.4e-158	0.0148445	0.3084	0.983922
rs36083532	5.8e-128	0.0133545	0.309436	0.976715
1:10831584 GA G	2.9e-123	0.0134094	0.282089	0.982814
rs10864470	2.4e-121	0.0132065	0.287272	0.986914
rs58064215	4.5e-118	0.0129445	0.287592	1.0
rs12031981	2.8e-112	0.0129622	0.267738	0.986998
1:10834614 GCC GC	7.5e-111	0.0129204	0.268212	0.979854
rs7512641	4e-108	0.0129773	0.257316	0.975738
rs7519604	2e-107	0.0131041	0.24374	0.985784
rs10864469	2.1e-107	0.0131037	0.243998	0.984678
rs205472	6.7e-104	0.0127324	0.249033	0.987825
rs205474	1.2e-103	0.0127166	0.248639	0.989514
rs6540948	1.1e-92	0.0122214	0.239372	0.985248
rs2235699	8.7e-92	0.0118013	0.274219	0.959489
1:10831714 TTCC T	4.8e-91	0.0126179	0.215452	0.97415

It should be noted that these R^2 values are obtained on a significance-enriched subset of SNPs and are therefore expectedly high; they should be interpreted as relative model comparison results, not absolute measures of genome-wide predictive accuracy.

Performance Comparison of ML Models

Table 1 presents the performance metrics of various machine learning models used in the study. As shown, the ETR achieved the lowest MSE and MAE, with the highest R^2 value, indicating its excellent predictive performance. In contrast, the EN model underperforms significantly, with high error values and a near-zero R^2, suggesting poor fit. This detailed comparison aids in identifying the most effective model for the prediction task.

Where,

- MSE is measured in base pairs squared (bp^2).
- MAE is measured in base pairs (bp).
- R^2 is unitless.

It is important to note that the high R^2 values obtained, such as 0.9991 for Extra Trees Regressor, are a direct consequence of evaluating models only on the most significant SNPs. Since these SNPs are highly predictive, the models exhibit near-perfect performance on this dataset (Table 2).

We used *P*-values to rank SNPs in order of statistical significance, with lower values denoting a higher connection with the trait. The SNP with the highest ranking is probably the one with the greatest biological significance. To make sure that only statistically significant relationships are taken into account, SNPs are ranked according to their *P*-value. *P*-values show the degree of confidence in the link, whereas effect size indicates the biological impact.

We compared our top-ranked[11] variations (based on *P*-values and feature importance) with known trait-associated loci documented in earlier GWAS in order to assess the biological significance of the discovered SNPs. Our investigation found that a number of important SNPs coincide with known loci linked to Varicose veins.

Conclusion

Experimental results indicate that ensemble models, especially the Random Forest and Extra Trees Regressors, perform optimally for SNP significance prediction, showing high accuracy and low error. These models leverage ensemble techniques, which significantly improve their ability to capture complex data relationships, achieving top scores in metrics like R-Squared, Explained Variance, and MAE. Models of gradient boosting like XGBoost and CatBoost also perform robustly, though with slightly higher RMSE and MSE. In contrast, linear models, particularly EN, demonstrated limited performance, struggling to capture genetic data complexity effectively. This analysis highlights the suitability of ensemble learning techniques for SNP significance prediction tasks. Future research could build upon these findings by refining model parameters and exploring hybrid models to address the unique challenges posed by large genetic datasets. Because model development and evaluation were performed on a subset of the most significant SNPs, the reported metrics reflect predictive performance within this enriched segment rather than across

the entire genomic landscape. This was a deliberate design decision to enable computationally tractable, exploratory model benchmarking. Broader validation strategies, such as chromosome-wise cross-validation or random subsampling, are planned for future work. Although advanced validation procedures such as cross-validation and hyperparameter tuning were not applied in this exploratory study, the consistent set-up provides a fair baseline for model comparison. Future work will extend this framework with rigorous validation strategies to evaluate generalizability more thoroughly.

References

1 Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15(6):335–46. http://doi.org/10.1038/nrg3706

2 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785

3 Breiman L. Random forests. Mach Learn. 2001;45:5–32. https://doi.org/10.1023/A:1010933404324

4 Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems; 2017. p. 3149–57.

5 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B StatMethodol. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x

6 Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory; 1992. p. 144–52.

7 McCarthy MI, Abecasis GR, Cardon LR, et al. Genomewide association studies for complex traits: consensus, uncertainty, and challenges. Nat Rev Genet. 2008;9(5):356–69. http://doi.org/10.1038/nrg2344

8 Chen Q, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99(6):323–9. http://doi.org/10.1016/j.ygeno.2012.04.003

9 Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems; 2018. p. 6638–48.

10 Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714–21. http://doi.org/10.1093/bioinformatics/btp041

11 Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. 2010;11(1):49. http://doi.org/10.1186/1471-2156-11-49

12 Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for Random Survival Forests using a conditional inference framework. Stat Med. 2019;36(8):1272–89. http://doi.org/10.1002/sim.7212

13 Zhang Z, Ersoz E., Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2020;42(4):355–60. http://doi.org/10.1038/ng.546

14 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucl Acids Res. 2012;38(16):e164. http://doi.org/10.1093/nar/gkq603

15 Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1

16 Smola AJ, Scholkopf B. A tutorial on support vector regression. Stat Comput. 2004;14(3):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88