Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing POLB Mutation Classification with a Random Forest and PSO Hybrid Model

Yasir Hussein Shakir[1], Doaa Yaseen Khudhur[2], Eshaq Aziz Awadh AL Mandhari [3], Ali Alkhazraji[4]
[1]*College of Graduate Studies (COGS), Universiti Tenaga Nasional (UNITEN), Kajang, Malaysia.*
[2]*College of Computer Science and Information Technology, University of Anbar in Iraq.*
[3]*Graduate School of Technology at Asia Pacific University of Technology and Innovation (APU) in Malaysia.*
[4]*Computer Science Department - Faculty of Sciences, Lebanese University, Hadat Campus, Beirut, Lebanon.*

ARTICLE INFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Cancer formation and development are mainly attributable to DNA damage. When DNA is damaged, the rate of genetic mutations increases, hence the need for DNA repair mechanisms. Another essential element that should be further considered while discussing the base excision repair and its impact on the maintenance of genome stability is DNA polymerase β (Polβ), encoded by the POLB gene. This enzyme is used in humans to fix damaged DNA strings. The purpose of this study is to develop an accurate risk predictive model for cancer associated with a specific mutation. Hybrid Machine Learning (HML) classification algorithm has been applied to the POLB SNPs dataset. Random Forest combined with Particle Swarm Optimization (PSO) algorithm's hyperparameters to find and extract the best parameters. Through the models, the RF-PSO demonstrated superior performance, achieving an accuracy of 84.06%, precision of 84.49 %, sensitivity of 84.06%, specificity of 90.55%, and an F1 score of 83.81%.To verify the performance of the proposed algorithm, the accuracy of the suggested RF-PSO classifier model was compared with another state-of-the-art model classifier, Naive Bayes, K Nearest Neighbors, Stochastic Gradient Descent, Linear Discriminant Analysis, Gradient Boosting Machines, AdaBoost, Passive Aggressive, Extra Trees, and Hist Gradient Boosting. The results also proved the superior ability of the implemented RF-PSO model classifier in the classification to investigate the relationship between POLB gene variations and their potential role in cancer onset. providing a robust foundation for further clinical applications and which will further help in better cancer diagnosis and treatment of the disease.

Keywords: Machine Learning; Classification; DNA Damage Repair; POLB Mutation; Random Forest; PSO.

INTRODUCTION

Cancer is a multifactorial disease driven by genetic mutations and DNA damage, which are well documented to be instrumental in cancer development [1]. The American Cancer Society estimates the number of new cancer cases and cancer deaths each year in the United States and offers the most current information about population-based cancer incidence and mortality by using data from the National Cancer Institute's central cancer registries up to 2020 and mortality data collected by the National Center for Health Statistics up to 2021. According to the data indicated by International Agency for Research on Cancer, updated for the year 2024 reveal that 2,001,140 new cases of cancer and 611,720 cancer related deaths will occur in the United States [2]. Clearly, DNA damage builds over time and can cause a higher risk of mutations and thus, require operational DNA repair processes to prevent damage build-up [3]. Among these repairs mechanistic modelling of DNA repair for double stand breaks (DSB), single strand breaks (SSB) and base damage (BD), shows the complexity of DNA damage is responsible for the longer repair times and the reason for the biphasic feature of mammalian cell repair curves [4]. A study by Sawyer DL [5] has revealed that the residues involved in coordination of the? -phosphate group of the dNTP are critical in determining nucleotide selectivity, polymerase activity and a fidelity. This study also draws attention to future investigation of this novel human POLB variant in vivo. Since Pol? is responsible for repairing DNA, any genetic variations of POLB may interfere with repair and enable mutation which can promote cancer development. DNA repair enzyme genes are known to include several

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SNPs, some of which have a significant impact on their activity [6,7]. In some cases, SNPs can decrease the efficiency of DNA repair, accumulate genomic mutations, and increase the vulnerability of an individual to developing cancer [8].

This paper offers an intelligent hybrid machine learning model to enhance cancer onset by developing a risk prediction model. Its main aim is to investigate the relationship between POLB gene variations and their role in cancer onset by developing a risk prediction model. The contributions of the paper are:

1. Propose a new learning algorithm (labeled RF-PSO) that classifies cancer associated with a specific mutation.
2. Preprocess, analyze and normalize the POLB gene variations dataset.
3. To utilize Random Forest for ranking features related to cancer-associated mutations.
4. Prove experimentally that proposed model outperforms other Intelligent models minimize the negative weighted F1 score.
5. Examine ten machine learning models on POLB gene variations and then demonstrate the most effective detector through experiments.

The subsequent sections of the paper are structured as follows: Section 2 review of the literature, review of the literature, followed by the methodology in Section 3.

In Section 4, the performance values were presented, and the results were compared. The section 5, discussion. And the final section 6 conclusion.

LITERATURE REVIEW

Cancer is a multifaceted pathology that relies on inheritable changes in the genetic structure of an organism that led to disruption of optimal performance of the various units of its cells. In addition, the application of machine learning (ML) in oncology has a great potential to increase the effectiveness of diagnostics and treatment of complex diseases including Cancer of Unknown Primary (CUP). The conventional ML algorithms, fine-tuned by large datasets on molecular analysis, can identify the primary site of metastatic eccrine tumors with significant precision. It has been reported that genomic profiling driver mutations and CNVs integrated into the diagnostic process of CUP and offer valuable information about tumor-specific oncogenic changes and Mutational signatures (MS) have also been described as useful in the diagnosis of CUP, since there are known signature MS associated with certain causes such as UV light induced DNA damage and tobacco [9].In the case of PC, somatic mutations on POLB gene are considered to affect the enzyme to its functionality and hinder the biological process of microsatellite; instability, loss of heterozygosity which result to the advancement of prostate cancer. Similarly, the studies conducted on the POLB gene concerning the prostate cancer show that they have over (20) mutations most of which are found on more than (50%) of the tumor chromosomes proving their importance in cancer development [10]. Valencia et. al [11], the LocalFilterNet (LFNet) is used for localizing protein-coding potential combined with enhanced classification by learning translational RNA to proteins. KafiKang et. al., [12] in a study to analyze turkey reovirus variants was able to use clustering methods to distinguish new variant after the other and was therefore able to demonstrate the usefulness of machine learning in identifying these emerging variants. In Black et. al,[13] the authors discussed the use of hyperspectral imaging in tumor classification problems and showed that machine learning can be used in discriminating between different types of tumors. Lorkowski et. al., [14] also highlighted the role of artificial intelligence and machine learning in diagnosis and management of Cancer of Unknown Primary (CUP); the authors underlined the changes that these technologies bring to the approach of difficult diagnostics. Tamehisa et. al., [15] also developed models of subtypes of uterine leiomyoma under the umbrella of machine learning and applied support vector classification and logistic regression in diagnosis. Zuo et al., [16] stated that the evaluation of the predictive model's performance should involve cross-center data validation and pointed out that the visualization and interpretation of these models is imperative to machine learning. In the work of Zhao et. al., [17] a transformer network for EGFR classification was proposed under the name of GMILT that combines multi-instance learning and discriminative weakly supervised feature learning for enhanced predictive capability. In a study by Nakagaki et. al., [18], a deep learning-based approach was adopted for prediction of IDH1 gene mutation using histopathological imaging and clinical data and established the possibility of using machine learning in genetic mutation prediction. Endometrial cancer patients have been identified using stemness-related signatures. Pang et. al, [19] discuss the

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

concepts of machine learning algorithm subtype risk modeling for discriminating between the prognosis, immune response, and somatic mutation. Davies et. al. [20] used molecular dynamics and machine learning for mutational hotspots classification. The combined strategy of random forest classification and feature selection highlights that the topological features are useful for the classification of adduct sites. Currently, the literature review indicates that machine learning and bioinformatics significantly contribute to the analysis of changes in POLB associated with cancer. These tools offer important information on the cancer and may be helpful for identifying potential targets for treatment. Marchevsky et al. [22] studied that, with the help of LDA models, 62.5% to 87.5% cases have been classified perfectly. Cohen's Kappa values for the compared classifications given by all models ranged from 0.25 to 1.0. They opined that by adopting artificial neural network, (ANN) classifications employing DNA methylation signature of SCLC and NSCLC cell lines possessed substantial to perfect concordance. However, as illustrated in the last entry of Table 4, LDA models had only poor to substantial agreement. Their work backs the possibility of using ANN analysis of DNA methylation data in the creation of automated lung cancer classification models.

Recently, Alkhanbouli et al. [23] focused on the effect of POLB gene product, DNA polymerase? (Pol?), involved in DNA repair and cancer. The mutations in POLB may lead to different DNA repair mechanisms, including Single Nucleotide Polymorphisms (SNPs). SNP sequences were analyzed bioinformatically to obtain extract the features while eight risk models were built using machine learning algorithms to identify cancer risk due to the mutations. XGBoost, Random Forest, and Weighted random forest turned out to provide the best ensemble solutions with the accuracy of 82%. The research highlights the role of POLB gene polymorphisms and reveals the usefulness of machine learning-based approaches in genomic cancer investigations.

MATERIALS AND METHODS

The three-stage approach suggested for the study of the potential association between genetic polymorphisms in the POLB gene and cancer development is revealed. The nature and design of this methodology in terms of the flow of the different steps are shown in the following Figure 1. Below are the details of each stage:

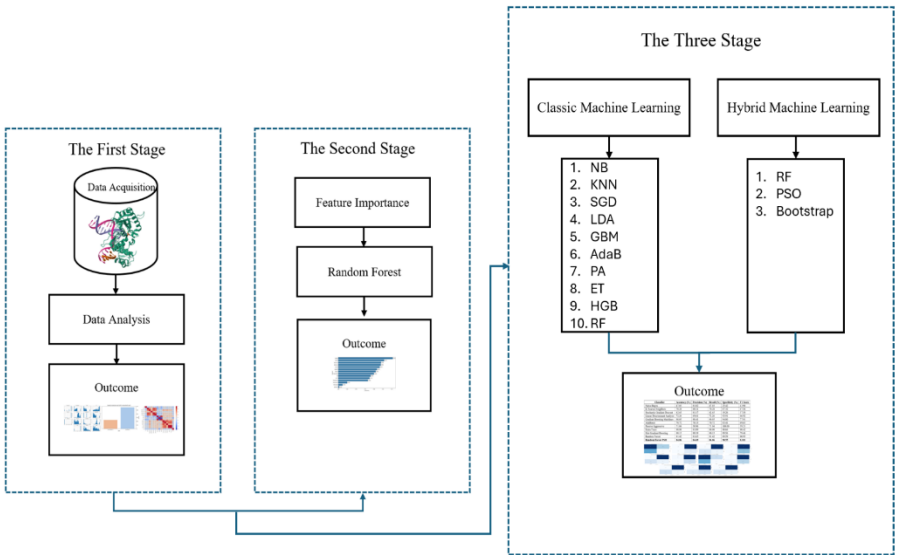

***Figure 1**: Proposed Methodology Model*

3.1. The First Stage

3.1.1. Data Acquisition

We used the dataset available in the Mendeley Data repository at [24], including 15 features and 813 samples or instances. The SNPs and their bioinformatics attributes associated with the POLB gene. These specific SNPs are associated with gene alterations whose occurrence is related to the development of cancer. Table 1 shown the dataset description of the POLB gene dataset.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

***Table 1**: Dataset Description.*

Description	Count
Number of Features	15
Sample Count	813

Table 2 shows 15 features, consisting of integers, floats, and descriptions.

***Table 2**: Data sample description types.*

No.	Feature Name	Datatype	Description
1#	PP	int64	Prediction score for protein impact (possibly from PolyPhen).
2#	SIFTR	float64	Score predicting the tolerance or intolerance of a mutation (SIFT).
3#	Polyphen2R	float64	PolyPhen-2 score assessing the damaging potential of mutations (Radical category).
4#	Poly1hen2P	int64	Likely a variant of PolyPhen-2 scores (Potential category).
5#	PROVEANR	float64	PROVEAN score indicating the mutation's deleterious potential (Risk category).
6#	PROVEANP	int64	PROVEAN score indicating the mutation's neutral or positive impact.
7#	CADDS	float64	CADD score estimating the deleteriousness of a mutation.
8#	CADDR	float64	CADD score with reference data for assessing mutation impact.
9#	CADDP	float64	CADD score with population data for predicting mutation effects.
10#	fathmmS	float64	FATHMM score predicting the functional impact of a mutation (S category).
11#	fathmmR	float64	FATHMM score predicting mutation effects (Risk category).
12#	fathmmP	int64	FATHMM score predicting mutation impact (Potential category).
13#	phyloP	float64	Phylogenetic conservation score, indicating evolutionary importance.
14#	phyloPR	float64	PhyloP score for a specific region (Risk category).
15#	class	int64	Target class indicating whether the mutation is deleterious (1) or neutral (0).

3.1.2. Data Analysis

This stage focuses on identifying the SNPs are linked to gene variations associated with cancer onset characteristics that enable the research's desired expectations. Ensuring representativeness is the main target of this stage. For each feature, statistical functions have been performed to calculate minimum, maximum, mean, 25%, 50% and 75%, and standard deviation (as computed in Appendix A). A histogram has been plotted for visualization. Instructions were used to produce a heatmap of the correlation matrix that was determined. Every cell contains the value of a correlation coefficient of two characteristics. The coefficient values range from -1 to 1 with -1 basing a perfect negative association with 1 basing a perfect positive association and 0 basing no association at all.

3.2. The Second Stage

3.2.1. Feature Importance Random Forest

In this stage, the Random Forest algorithm is used for feature importance analysis, In Random Forest, feature importance is computed based on how much a particular feature contributes to reducing the impurity (Gini impurity or entropy) across all the trees in the forest. Specifically, feature importance for a feature f is calculated as the sum of the decrease in node impurity across all trees in the forest, weighted by the number of samples that reach the node as shown by the following equation:

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$I(f) = \sum_{i=1}^{T} \sum_{n \in node(t)} 1\{split\ no\ f\} \frac{N_n}{N} \Delta i_n \quad (1)$$

3.3. The Three Stage

In this stage, we implemented a variety of machine learning algorithms and divided them into classic machine learning and hybrid machine learning

3.3.1. Classic machine learning

They are now passed on to the classical machine learning classifiers to train an ML model that could classify cancer associated with a specific mutation as deleterious or neutral. Brief information about the machine learning classifiers considered for the research is discussed [21].

Naive Bayes (NB):

A model integrating a probabilistic classifier following Bayes' Theorem and supposing that features are independent. Indeed, it is easy to implement, fast and efficient particularly for text classification.

K-Nearest Neighbors (KNN):

A global instance-based learning algorithm that forms the basis of many algorithms. It classifies the result from data points by voting as informed by the k nearest training samples in the feature space. It works well especially in low dimensions but when dealing with big data or data with many features, it becomes slow and least accurate.

Stochastic Gradient Descent (SGD):

Algorithms utilized to arrive at a minimum of a loss function accomplished by subsequent adjustments of the size of a weight vector, particularly relevant to massive ML computations. Res ranges from fast and effective to slow and ineffective depending on hyperparameters such as the learning rate.

Linear Discriminant Analysis (LDA):

A technique that requires that the values of the target variables from each class are normally distributed. It throws data into lower-dimensional space where the discrimination between classes is at its lowest. It is used frequently in dimensionality reduction and classification process.

Gradient Boosting Machines (GBM):

A capable means of constructing multiple decision trees one after another in a variety of sequences. One tree adjusts for mistakes made by the previous tree, and so on, to enhance the accuracy. While GBMs are useful for structured or tabular data they are not always the fastest to train.

AdaBoost (AdaB):

A boosting algorithm that makes a strong classifier from a few relatively weak classifiers, which are commonly but not necessarily decision trees. It focuses on training on examples that it got wrong in the past trying to make it learn better. Suitable for the treatment of datasets with a limited number of observations on the data and a large number on the classes.

Passive Aggressive (PA):

An evaluation methodology or algorithm for learning in large data using Internet or World Wide Web environment. It trains it only when it makes wrong prediction or prediction of type aggressive and otherwise leaves the model the same or makes a prediction of type passive. Fairly often applied to text categorization problems.

Extra Trees (ET):

An instance of the ensemble method, which resembles Random Forests, but which has a different way of selecting splits in trees. It splits the features randomly which makes the model train faster and more random yet very accurate for classification and regression.

Hist Gradient Boosting (HGB):

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A type of gradient boosting that starts by finding a correlation between the features and the target value before extending towards the production of models with higher or lower accuracy needed for big data sets. It employs the binding of continuous variables in histogram form leading to faster and more efficient memory usage as compared to the standard gradient boosting algorithms.

Random Forest (RF):

A popular technique that creates different decision trees and integrates the output of each of them (average for regression analysis and voting system for classification). First, it is built to be strong, it deals with the case when some values are missing, and finally, it minimizes the problem of overfitting as it gives an average of several trees.

3.3.2. Hybrid machine learning

This paper investigates the application of an innovative classification algorithm, the Random Forest with Particle Swarm Optimization. The RF technique was used to establish the modeling of the nonlinear mapping function between the input features and the target output. To enhance the RF's performance, Particle Swarm Optimization (PSO) was applied to tune two critical hyperparameters: The two tuning parameters include the number of estimators; that is the number of trees to be constructed (n_estimators) and the maximum depth of the trees to be constructed (max_depth). Three hyperparameters are noteworthy as they have major impacts on model accuracy and generalization capability. Ask those tuning them to make the right adjustments to obtain expert performance. Consequently, the mechanism of PSO optimization was developed to identify the least achievable value of the negative weighted F1 score, which represents the objective function. For each set of hyperparameters, an RF model was generated from a stratified training dataset and further tested on a stratified test dataset. The main goal was to achieve a high F1 score to balance the value of precision as well as recall in the last model. As a means of minimizing randomness in the effect of the RF, PSO was done using (100) particles for (50) iterations, while limiting the range of n_estimators between (10) and (500) and max_depth between (5) and (50). As a result of performing PSO, the best values attained for these hyperparameters were for n_estimators = X and max_depth = Y (insert actual numbers). After fine-tuning performances of the Random Forest model, Hyperparameters which gave the best optimization results were used to retrain the Random Forest model. To ensure robust validation of the model's performance, bootstrapping was conducted with (10,000) iterations. In each iteration, a random sample of the training data (with replacement) was used to retrain the RF model, and classifications were made on the test set. The average F1 score from each iterations were stored and is presented in the immediate work, then the average F1 score of each model was used as the final evaluation of the models performance. Analyzing the model this manner gave a fair assessment of the generalization ability and stability of the model. The PSO- optimized RF model fared well as against models with hyperparameters set at default thus validating the infallibility of the PSO in circumstances of hyperparameter vector optimization. Algorithm 1 presents the RF-PSO proposed classifier for mutations in the POLB gene. The proposed

Algorithm 1: RF-PSO Classifier for Mutations in the POLB Gene

Input: Dataset (D)
Output: accuracy, precision, sensitivity, specificity, and F1 score
Begin
Split (D, X_train, X_test, Y_train, Y_test) using stratified sampling;
Define the objective_function (params);
    n_estimators, max_depth ← int(params[0]), int(params[1]);
    if n_estimators < 1 or max_depth < 1 then return float('inf');
    Initialize RFmodel (rf) ← RF(n_estimators=n_estimators, max_depth=max_depth);
    rf.fit(X_train_strat, Y_train_strat) ;
    y_pred ← rf.predict(X_test_strat) ;

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

```
    Calculate fitness score ← -f1_score (Y_test_strat, y_pred, average='weighted');
return fitness score;
end objective_function ;
Define PSO parameter bounds (lb, ub) ; // lb = [10, 5], ub = [500, 50]
best_params, best_f1 ← pso (objective_function, lb, ub, swarmsize=15, maxiter=20,
minstep=1e-8, minfunc=1e-8);
best_n_estimators ← int(best_params[0]), best_max_depth ← int(best_params [1]) ;
rf_optimized ← RandomForestClassifier(n_estimators=best_n_estimators,
max_depth=best_max_depth, random_state=42) ;
rf_optimized.fit(X_train_strat, Y_train_strat) ;
y_pred_rf ← rf_optimized.predict(X_test_strat) ;
Initialize n_iterations ← 10000;
Initialize dictionaries (accuracy, precision, sensitivity, specificity, and F1 score);
for i ← 1 to n_iterations do
    Bootstrap sampling on (X_train_strat, Y_train_strat);
    Train the rf_optimized model on bootstrapped samples;
    y_pred_bootstrap ← rf_optimized. predict(X_test_strat);
    Update accuracy, precision, recall, f1score, TNR_dict, TPR_dict for the current iteration;
end for i;
Calculate the final Measurement ← accuracy, precision, sensitivity, specificity, and F1
score (Y_test_strat, y_pred_rf);
Return accuracy, precision, sensitivity, specificity, and F1 score ;
End FR-PSO
```

pseudocode is given in (Table 3).

3.3.3. Performance Evaluation Metrics

The performance of the machine learning models was assessed using several evaluation metrics and these five prominent class measures are defined (Accuracy, Precision, Recall, Specificity, and F1 Score). These metrics gave information on the ability of each model to accurately classify the cancer related mutations according to the POLB SNPs dataset.

1- Accuracy: Determines the completion accuracy of the model and represents the ratio between the numbers of true results (both in the positive and in the negative classes) and the total number of cases, by the following equation.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (2)$$

2- Precision: Shows the proportion of the correct positive results concerning all the positive results. That is, a high level of accuracy means that few erroneous readings are detected, by the following equation.

$$\text{Precision} = \frac{TP}{TP+FP} \times 100 \quad (3)$$

3- Recall: It is also known as sensitivity and indicates the proportion of actual positive cases that have been recommended by the model, by the following equation.

$$\text{Recall} = \frac{TP}{TP+FN} \times 100 \quad (4)$$

4- Specificity: The model must accurately exclude cases that are non-cancerous, this is represented by the true negative parameter, by the following equation.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 \quad (5)$$

5- F1 Score: The f-measure, considering both precision and recall at the same time and being ideal for use with non – balanced datasets, by the following equation.

$$\text{F1 Score} = 2 \times = \frac{Precision \times Recall}{Precision + Recall} \times 100 \quad (6)$$

RESULTS

Precisely, we evaluated the performance of the proposed model on the cancer associated with a specific mutation testing dataset. Thereafter, we tested four fundamental classic machine learning algorithms (NB, KNN, SGD, LDA, GBM, AdaBoost, PA, HGB and RF) on the same dataset and compared results. Furthermore, we examined our model against hybrid machine learning models (RRFPSO). To attach theoretical proposals into concrete outcomes and code patterns, we use the Python machine learning libraries: sklearn, pandas, matplotlib, and seaborn. All algorithms were implemented in Python with Pandas 2.1.4, Scikit-learn 1.5.2, Matplotlib 3.7.1, and Seaborn 0.13.1 frameworks. Codes were executed on Google Colab with the support of the CPU Tesla T4 2000.148 MHz and 15GB RAM. The dataset's analysis and visualization findings are first shown. The outcomes of the proposal's execution and associated models are then provided. All the pretrained models are accessible on GitHub:

https://github.com/yasserhessein/Cancer-associated-mutations-of-POLB.

4.1. Data Analysis Outcome

The dataset, which contains 813 samples, and 15 characteristics linked with POLB gene SNPs, was preprocessed to assure quality and representativeness. The statistical analysis of each characteristic, as shown in Appendix A, gave important insights, and histograms were utilized to display the distribution of the features. The figure 3 displays the class Distribution in the POLB Gene Variations Dataset, the number of neutral (class 0) 577 instance and deleterious (class 1) gene variations 236 instances. Another fact is that the distribution of the dataset is highly skewed with most of the data samples labelled as (class 0) neutral variations.

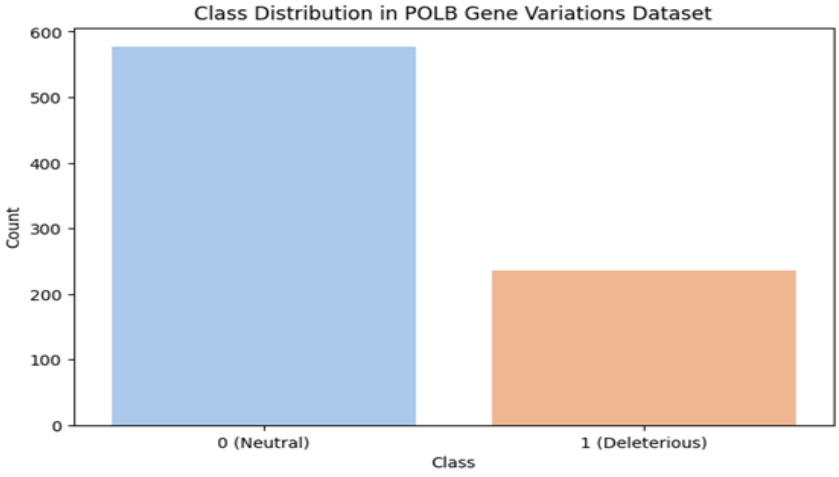

Figure 4: Class Distribution in the POLB Gene.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

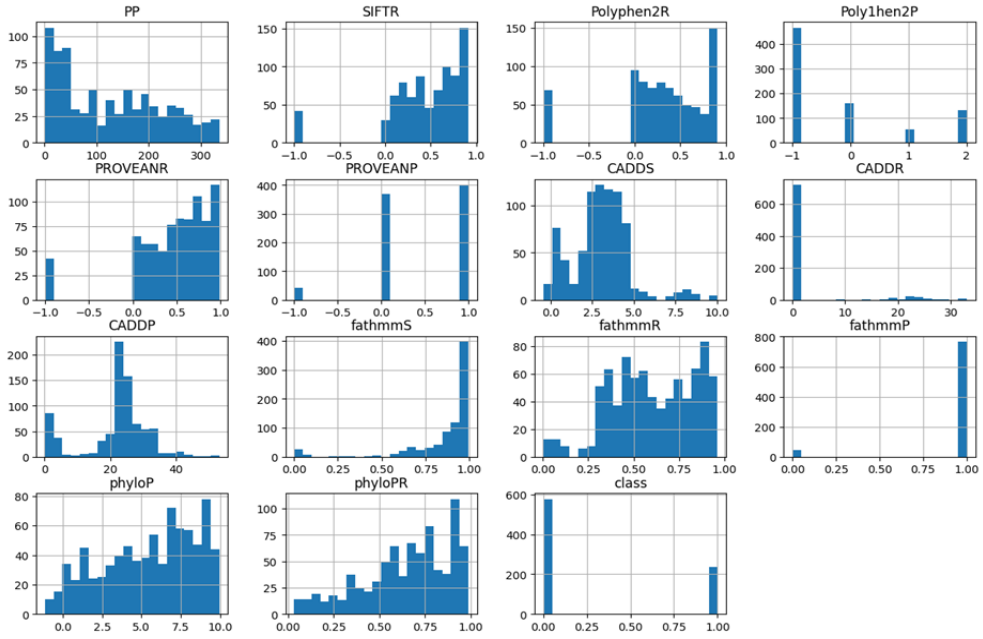

Figure 5: A histogram of POLB Gene Features.

The histograms provide an overview of the distribution of each scaled feature. Most features have relatively normal distributions, though a few exhibits skewness as shown in figure 4. Most of these features are skewed towards specific value ranges (SIFTR, Polyphen2R, and PROVEANP show a clustering of values around 0.5 to 1.0). PP shows a long tail, with values spread across a wide range, indicating potential outliers or a non-uniform distribution. CADDS appears to be more normally distributed, whereas CADDR is highly skewed with many instances clustered at specific values (around 0, 1, and 2). CADDP shows some spread, but with a peak towards certain values, suggesting less variation and fathmmS and fathmmR have a skewed distribution with more values concentrated at the higher end of the scale. fathmmP shows a significant skew, with many instances around 0 and a few outliers at higher values.

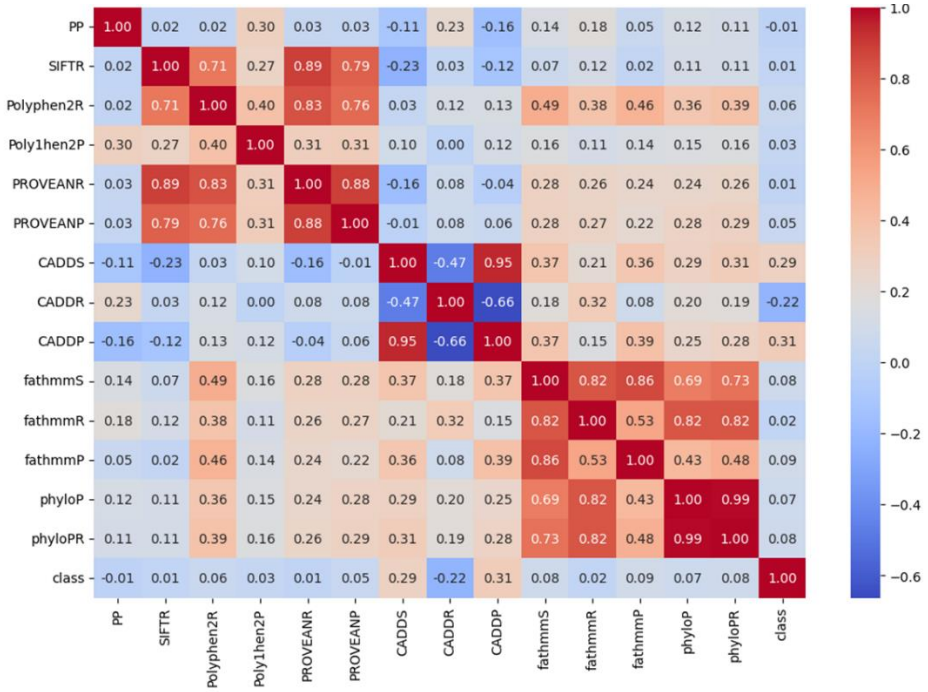

Figure 6: Features Heatmap of POLB Gene

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The heatmap you shared presents Feature Correlation Matrix various features show figure 5 and Correlations between features range from -1 to 1 and where values closer to 1 indicate a strong positive correlation and values closer to -1 indicate the strong negative correlation. CADDS and CADDP are highly positively correlated with the coefficient of 0.95 indicating redundancy between these features. These two seem to be very closely related and could therefore be dropped or combined if one wants to minimize multicollinearity in the model. Furthermore, there is increased correlation (0.86) between CADDS and fathmmS which also analyze similar patterns from the data set. CADDS and CADDR are also negatively related with coefficient of (-0.47) implying that they have an opposite direction. These opposing tendencies may signify different patterns in the dataset PhyloP and phyloPR correlate almost perfectly positively (0.99) which means that PhyloP and phyloPR are almost identical. To make one of these features a constant one - which frequently might be helpful to enhance the variance of additional machine learning features by excluding features that are highly correlated with each other - we found that the features fathmmS and fathmmP are considerably correlated with a correlation coefficient of 0.86, thus it might imply that they measure similar aspects of variation in the dataset. The problem of multicollinearity could occur if both are retained in the model. There is a moderate positive correlation between PROVEAN and Polyphen2 (0.83), which indicates they may do similar things: provide similar data. CADDP bears a moderate negative relationship with CADDR (-0.66), indicating that these two metrics have some inverse movement.

4.2. Feature Importance Random Forest Outcome

In this outcome stage, Gini impurity was utilized in the Random Forest (RF) to rank the feature importance. As show in figure 6, the top three features contributing to cancer-related mutations in the POLB gene were CADDS importance of 0.122, CADDP importance of 0.102, and fathmmR importance of 0.095. These features produced the best performers that also showed a high degree of purity implying their importance in this classification of deleterious mutations. Other notable features include fathmmS importance of 0.095, CADDR importance of 0.092, and phyloPR importance of 0.086, all contributing meaningfully to the model's performance.

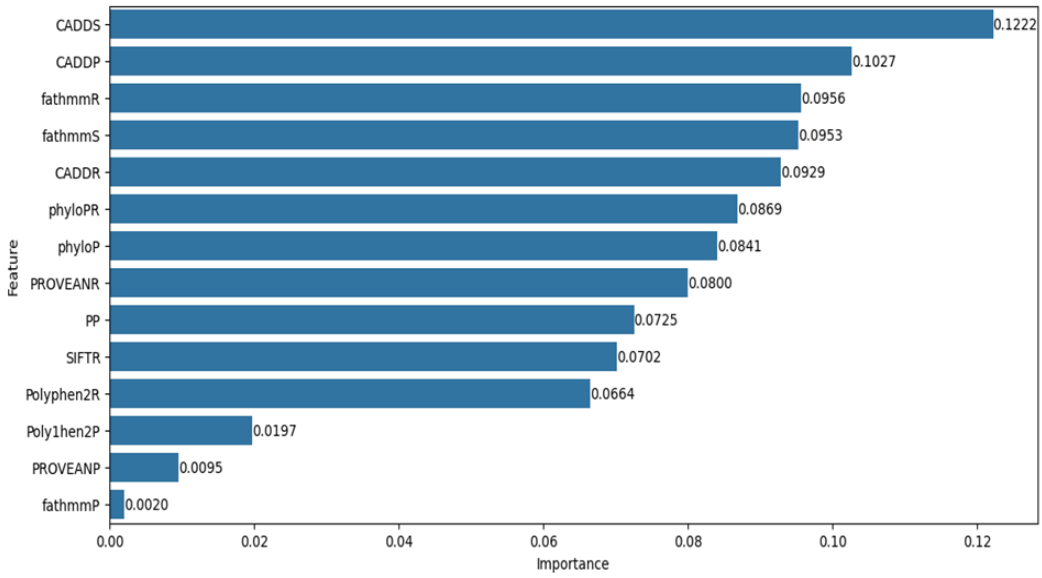

Figure 7: Feature Importance Outcome

4.3. Machine Learning Outcome

The machine learning models used for classification of POLB SNPs dataset were expected outcome of cancer risk attached with mutation in POLB gene as presented in the Table 3. In this paper, we have implemented several classical and hybrid machine learning algorithms, and their performance were compared since factors like accuracy, precision, recall, specificity and F1 score. The Random Forest with Particle Swarm Optimization RF-PSO hybrid model proved to be best among the classifiers having accuracy of 84.06%, precision of 84.49%, recall of 84.06%, specificity of 90.55% and F1-score of 83.81%. The outcompeting with the RF model has again demonstrated the

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

optimal performance of the RF-PSO model for feature selection and minimizing false negatives on which the cancer risk prediction is based. Similarly, Naive Bayes Classifier, K Nearest Neighbors, Gradient Boosting Machines, Ada Boost Machine, and few others performed different levels of wellness. Although, both GBM and Hist Gradient Boosting exhibited good accuracy percentages of 80.05% and 80.13%, respectively they lacked in the precision and recall of RF-PSO. This is because, according to the research carried out herein, the application of the PSO algorithm to the fine-tuning of the RF model yields a higher level of predictive accuracy in the identification of cancer related mutations. These improvements in performance demonstrate the effectiveness of the hybrid machine learning methods in more complicated genetic databases.

Table 3: Performance of Machine Learning Models

Classifier	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score
Naive Bayes	47.03	81.82	47.03	25.42	43.90
K Nearest Neighbors	70.28	68.11	70.28	87.32	67.56
Stochastic Gradient Descent	42.65	81.17	42.65	19.26	37.50
Linear Discriminant Analysis	72.26	69.61	72.26	93.91	66.92
Gradient Boosting Machines	80.05	80.41	80.05	94.80	77.91
AdaBoost	70.72	70.13	70.72	83.41	69.63
Passive Aggressive	71.04	50.96	71.04	100.00	59.21
Extra Trees	80.80	81.09	80.80	88.66	80.42
Hist Gradient Boosting	80.13	80.19	80.13	89.56	79.46
Random Forest	81.42	81.63	81.42	89.59	80.95
Random Forest PSO	**84.06**	**84.49**	**84.06**	**90.55**	**83.81**

The confusion matrices shown in the figures 5 reveal the performance of different machine learning models for classifying risk for cancer from POLB SNPs dataset. The best model is the RF-PSO providing an equal proportion of misclassified instances in terms of false negative and false positive, crucial for cancer risk prediction. Random Forest and Gradient Boosting Machines give relatively good accuracy but have higher false negative values. Hence, only three of the models committee all the positive instances but have several false positives; Passive Aggressive and Linear Discriminant Analysis fail to classify most positive instances and are therefore less preferable.

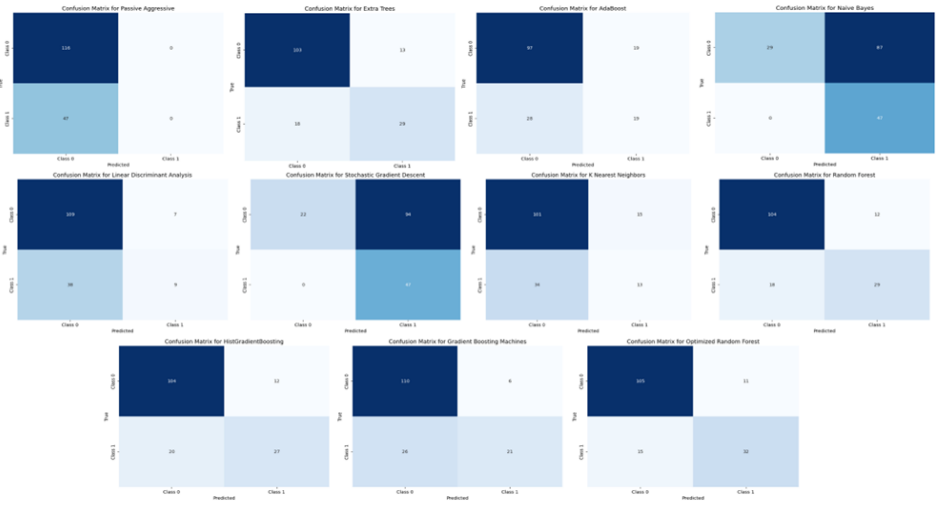

Figure 8: Confusion Matrices Performance all the Models

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

DISCUSSION

The results in this study highlight the impact of model selection and feature importance for identifying the risk of cancer for POLB gene mutations. Among the classical and hybrid machine learning models tried herein, the proposed (RF-PSO) hybrid model provides the overall highest accuracy, precision, specificity, recall and F1 score all at once. This result demonstrates the importance of combining biomarkers in improving the accuracy of sophisticated genomics data. RF with PSO fine-tuning was more helpful in reducing false negatives, which is crucial for performance of cancer risk prediction, because omission of deleterious variants is a detrimental activity. Surprisingly, the original model of RF was slightly less accurate than the RF-PSO hybrid that demonstrates that optimization approaches such as PSO can fine-tune the model. This optimization probably assisted in the feature selection process and in general led to better classification on the deleterious or neutral mutations. Other models include GBM and Hist GB that also did a reasonable job of classification but poor in precision and recall. They are therefore prone to misclassifying the instances. The dataset was skewed towards the first class, class 0 corresponded to the neutral instances which undermined the performance of NBs, and SGD generally had a low accuracy and F1 scores. This means that for simple models or models that have not incorporated an optimization method, the negative impact of imbalanced data will be felt and the confusion matrices showed that models like PA and LDA though capable of classifying many negative instances, completely missed the positive or deleterious mutations making such models unfit for this purpose. The RF Gini importance feature showed that important features in predicting cancerous mutations included CADDS, CADDP and fathmmR. These features exhibited high levels of purity and significantly contributed to model performance. However, high correlations among some features suggested redundancy, which could be addressed in future models to reduce multicollinearity and improve efficiency.

CONCLUSION

In conclusion, this study demonstrates the importance of both classical and hybrid machine learning approaches in the domain of cancer risk classification. The RF-PSO hybrid model stood out as the most reliable and accurate method for classifying mutations in the POLB gene. The findings emphasize that optimizing models through algorithms like PSO can enhance the accuracy and robustness of predictions, particularly when dealing with complex and imbalanced genetic datasets. Additionally, feature importance analysis revealed that specific features played a pivotal role in model performance, which offers valuable insights for future research on genetic mutation analysis. While the RF-PSO model delivered the best results, the performance of Gradient Boosting Machines and Random Forest was also noteworthy. Future work should focus on addressing class imbalance and exploring other optimization techniques to further refine predictive models. Moreover, reducing feature redundancy and leveraging more advanced feature selection methods may improve model performance. Overall, the application of hybrid machine learning models, particularly those fine-tuned with optimization techniques, offers a promising path forward in the field of cancer genetics, providing more accurate and reliable tools for risk classification.

Appendix

Table 4: POLB Gene Mutation Features Statistics

Measure	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15
Minimum	1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-0.43	0.02	0.17	0.00	0.00	0.00	-1.13	0.03	0.00
25%	36.00	0.23	0.10	-1.00	0.25	0.00	2.18	0.44	20.70	0.84	0.44	1.00	3.26	0.51	0.00
50%	117.00	0.53	0.34	-1.00	0.57	0.00	3.09	0.62	23.60	0.94	0.59	1.00	5.84	0.69	0.00
75%	206.00	0.72	0.67	0.00	0.77	1.00	3.98	0.90	26.80	0.98	0.82	1.00	7.85	0.85	1.00
Maximum	335.00	0.91	0.91	2.00	0.98	1.00	10.01	33.00	53.00	0.99	0.96	1.00	9.93	0.99	1.00
Mean	127.36	0.45	0.32	-0.18	0.47	0.44	3.05	3.17	21.98	0.86	0.60	0.94	5.43	0.65	0.29

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Standard Deviation	97.87	0.44	0.50	1.12	0.44	0.59	1.74	7.37	9.91	0.22	0.23	0.24	2.96	0.24	0.45

REFRENCES

[1] Wang M, Xie C. DNA damage repair and current therapeutic approaches in gastric cancer: A comprehensive review. Frontiers in Genetics. 2022 Aug 12;13:931866.
[2] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: a cancer journal for clinicians. 2024 Jan 1;74(1).
[3] Anderson CJ, Talmane L, Luft J, Connelly J, Nicholson MD, Verburg JC, Pich O, Campbell S, Giaisi M, Wei PC, Sundaram V. Strand-resolved mutagenicity of DNA damage and repair. Nature. 2024 Jun 12:1-8.
[4] Nikjoo H, Rahmanian S, Taleei R. Modelling DNA Damage-Repair and Beyond. Progress in Biophysics and Molecular Biology. 2024 May 14.
[5] Sawyer DL. The S180R Human Germline Variant of DNA Polymerase ? Exhibits Low Fidelity and the Potential to Drive Cancer Formation (Doctoral dissertation, The University of Arizona).
[6] Hall J, Marcel V, Bolin C, Fernet M, Tartier L, Vaslin L, Hainaut P. The associations of sequence variants in DNA-repair and cell-cycle genes with cancer risk: genotype-phenotype correlations.
[7] Starcevic D, Dalal S, Sweasy JB. Is there a link between DNA polymerase beta and cancer? Cell cycle. 2004 Aug 21;3(8):996-9.
[8]Kladova OA, Fedorova OS, Kuznetsov NA. The role of natural polymorphic variants of DNA polymerase? in DNA repair. International Journal of Molecular Sciences. 2022 Feb 21;23(4):2390.
[9] Lorkowski SW, Dermawan JK, Rubin BP. The practical utility of AI-assisted molecular profiling in the diagnosis and management of cancer of unknown primary: an updated review. Virchows Archiv. 2024 Feb;484(2):369-75.
[10] Makridakis NM, Caldas Ferraz LF, Reichardt JK. Genomic analysis of cancer tissue reveals that somatic mutations commonly occur in a specific motif. Human mutation. 2009 Jan;30(1):39-48.
[11] Valencia JD, Hendrix DA. Improving deep models of protein-coding potential with a Fourier-transform architecture and machine translation task. PLOS Computational Biology. 2023 Oct 12;19(10):e1011526.
[12] KafiKang M, Abeysiriwardana C, Singh VK, Young Koh C, Prichard J, Mor SK, Hendawi A. Analysis of Emerging Variants of Turkey Reovirus using Machine Learning. Briefings in Bioinformatics. 2024 May;25(3):bbae224.
[13] Black D, Byrne D, Walke A, Liu S, Di Ieva A, Kaneko S, Stummer W, Salcudean T, Suero Molina E. Towards machine learning-based quantitative hyperspectral image guidance for brain tumor resection. Communications Medicine. 2024 Jul 4;4(1):131.
[14] Lorkowski SW, Dermawan JK, Rubin BP. The practical utility of AI-assisted molecular profiling in the diagnosis and management of cancer of unknown primary: an updated review. Virchows Archiv. 2024 Feb;484(2):369-75.
[15] Tamehisa T, Sato S, Sakai T, Maekawa R, Tanabe M, Ito K, Sugino N. Establishment of noninvasive prediction models for the diagnosis of uterine leiomyoma subtypes. Obstetrics & Gynecology. 2022 May 5:10-97.
[16] Zuo Y, Liu L, Chang C, Yan H, Wang L, Sun D, Ruan M, Lei B, Xia X, Xie W, Song S. Value of multi?center 18F?FDG PET/CT radiomics in predicting EGFR mutation status in lung adenocarcinoma. Medical Physics. 2024 Jan 29.
[17] Zhao W, Chen W, Li G, Lei D, Yang J, Chen Y, Jiang Y, Wu J, Ni B, Sun Y, Wang S. GMILT: a novel transformer network that can noninvasively predict EGFR mutation status. IEEE Transactions on Neural Networks and Learning Systems. 2022 Jul 21.
[18] Nakagaki R, Debsarkar SS, Kawanaka H, Aronow BJ, Prasath VS. Deep learning-based IDH1 gene mutation prediction using histopathological imaging and clinical data. Computers in Biology and Medicine. 2024 Sep 1;179:108902.
[19] Pang X, Wang Y, Zhang Q, Qian S. A stemness-based signature with inspiring indications in discriminating the prognosis, immune response, and somatic mutation of endometrial cancer patients revealed by machine learning. Aging (Albany NY). 2024 Jul 7;16(14):11248.

Journal of Information Systems Engineering and Management
2025, 10(46s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[20] Davies JG, Menzies GE. Utilising biological experimental data and molecular dynamics for the classification of mutational hotspots through machine learning. Bioinformatics Advances. 2024 Aug 26:vbae125.
[21] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007 Jun 10;160(1):3-24.
[22] Marchevsky AM, Tsou JA, Laird-Offringa IA. Classification of individual lung cancer cell lines based on DNA methylation markers: use of linear discriminant analysis and artificial neural networks. The Journal of Molecular Diagnostics. 2004 Feb 1;6(1):28-36.
[23] Alkhanbouli R, Al-Aamri A, Maalouf M, Taha K, Henschel A, Homouz D. Analysis of cancer-associated mutations of POLB using machine learning and bioinformatics. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2024 May 1.
[24] Alkhanbouli R, Alaamri A, Maalouf M, Taha K, Henschel A, Homouz D. POLB SNPs Dataset. Mendeley Data. 2024; V1. Available from: https://doi.org/10.17632/d6385g6kv6.1