ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

Machine Learning Based Classification of POLB Gene Mutations with SHAP Based Interpretability

Baratam Sai Lahari[1]; Balasani Manasini[2]; Rachapudi Reshma[3]; Sri. Ch. Ratna Babu[4]

Department of Computer Science and Engineering
R. V. R & J. C College of Engineering
Chowdavaram, Guntur, Andhra Pradesh, India

Publication Date: 2026/04/28

Abstract: This research focuses on developing a machine learning method to study mutations that are related to the development of cancer in the POLB gene based on the features associated with single nucleotide polymorphisms(SNPs). Initially, a dataset made up of bioinformatics-derived features like SIFT, PolyPhen2, CADD, and REVEL was pre-processed and subsequently used as a foundation for the creation of predictive models. Five types of classification algorithms were applied and assessed: Logistic Regression, Random Forest, Support Vector Machine, Multilayer Perceptron and XGBoost. To ensure that performance estimates were valid, bootstrap resampling techniques were employed and metrics including accuracy, precision, recall, F1 score and specificity were calculated. Results from the experiments showed that both ensemble models (Random Forest and XGBoost) produced the most accurate results approximately 83 percent which indicated that these models can capture complex relations in SNP data. In addition, SHAP explanation methods were used to explain model predictions and determine the features that had the largest effects on classification decisions. The study indicated that machine learning techniques have many applications in genomic research, particularly when it comes to outcomes associated with mutations that lead to cancers.

Keywords: *POLB, Single Nucleotide Polymorphism, Machine Learning, Cancer Mutation Prediction, Random Forest, XGBoost, SHAP Explainability.*

How to Cite: Baratam Sai Lahari; Balasani Manasini; Rachapudi Reshma; Sri. Ch. Ratna Babu (2026) Machine Learning Based Classification of POLB Gene Mutations with SHAP Based Interpretability. *International Journal of Innovative Science and Research Technology*, 11(4), 2227-2233. https://doi.org/10.38124/ijisrt/26apr1516

I. INTRODUCTION

A variety of genetic mutations that affect the functioning of cells can result in the development of cancer as one of the leading causes of death in the world. Genomic stability (the maintenance of the integrity of the genome) depends upon several biological pathways; however, one pathway that serves a very important purpose is the DNA repair pathway. DNA repair pathways fix damaged DNA to avoid the accumulation of mutations; thus, functioning in a manner to maintain genomic stability.

Base excision repair (BER) is a common DNA repair pathway. The POLB gene provides the genetic information to produce DNA polymerase beta (DNApolβ), the enzyme responsible for repairing segments of DNA that are damaged. When the POLB gene has mutations, the ability of the DNA repair pathway to repair DNA is diminished leading to genomic instability and the increased potential for the development of cancer.

Genetic variation within the human genome is primarily found in a single nucleotide polymorphism (SNP). SNPs are commonly present in the genome and occur as single base changes that may have an impact on the resulting protein's structure and/or function. Within the POLB gene, some SNPs may impact the efficiency of the DNA repair mechanism, thus contributing to the development and progression of cancer. Therefore, the identification and analysis of these genetic mutations are critical in establishing the mechanisms of disease as well as providing a means for early diagnosis.

Machine learning methodologies have gained attention in the field of bioinformatics due to their capacity for analysing large quantities of complex, high-dimensional data (i.e., biological). As such, these methodologies allow for the identification of patterns or relationships among the biological data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

II. LITERATURE REVIEW

Various studies have demonstrated the importance of the POLB gene and its function in DNA repair as it pertains to maintaining or restoring genetic stability. Furthermore, specific mutations that have occurred in the POLB gene have been linked with a number of different types of cancer, and therefore mutation(s) in this gene may negatively impact how effective base excision repair pathways are for that individual.

Multiple studies conducted over the past two decades have attempted to identify and analyse specific mutations of the POLB gene in order to understand how these mutations can contribute to the development of different types of cancer.

Single nucleotide polymorphisms (SNPs) are commonly examined as significant sources of genetic variation that can affect the function of proteins and influence an individual's risk for various diseases. Bioinformatics tools such as SIFT, PolyPhen2, Combined Annotation Dependent Depletion (CADD) and REVEL are generally used for predicting the potential impact that SNP(s) may have on the structure and/or function of any subjected protein. Each of these tools generates a feature-generated score for use in computational analyses after the initial SNP prediction has been completed.

In recent years, machine learning techniques have increasingly been applied to the field of bioinformatics regarding classifying mutations and predicting disease occurrence. Logistic regression, support vector machines, random forests, and gradient boosting algorithms are just some of the examples of models that have been used successfully for identifying patterns in the genomic data from the mutations in these types of studies. Each of these different approaches has been able to classify mutations according to their predicted likelihood of being associated with the development of particular diseases (e.g. cancer). Whereas prior studies have generally focused on using only one prediction method or bioinformatics tool, there is now a greater need for developers to use multiple algorithms, bioinformatics tools, or a combination thereof.

III. METHODOLOGY

The research focuses on a machine learning method for analyzing mutations linked to cancer specifically associated with the POLB gene using single-nucleotide polymorphisms (SNP) features. The method will address each stage of data handling including dataset use; data preprocessing; model creation and evaluation (using bootstrap resampling).

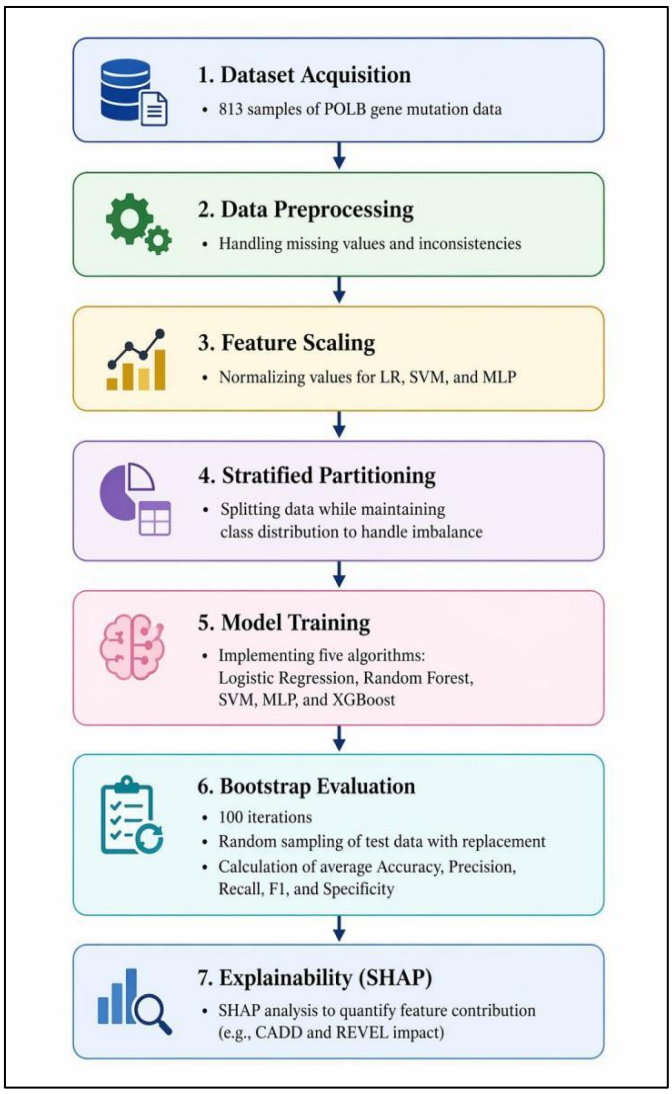

Fig. 1 Proposed Methodology Flow for POLB Gene Mutation Classification

Figure 1 illustrates how the proposed method for classifying POLB gene mutations works as a whole. It shows the steps that come after preparing and preprocessing the dataset, training the model, testing it, and making it explainable using SHAP.

A. Dataset

This research study utilized a data set of preprocessed SNP feature data generated from bioinformatics prediction tools. It contained 813 samples with 14 features for each sample, which are scores generated from tools like SIFT, PolyPhen2, Combined Annotation Dependent Depletion, and REVEL — scores generated from bioinformatic prediction tools (SIFT, PolyPhen2, Combined Annotation Dependent Depletion [CADD] and REVEL) were used as features for the formation of the dataset. Each sample was representative of one mutation and included the associated features and class labels indicating whether or not the mutation is likely associated with cancer. There is an imbalance between the two class distributions for mutations associated with cancer and those not associated with cancer, and this was considered when evaluating the model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

B. Data Preprocessing

To guarantee that the model receives consistent and reliable inputs, preprocessing was completed prior to using the dataset to train the model. During the preprocessing, missing values and inconsistencies were detected within the dataset and corrections were made where required. As the dataset contains scores from bioinformatics tools that are numeric features, it can be used directly in machine learning after completing preprocessing.

Models that are sensitive to the magnitude of features (e.g., Logistic Regression, Support Vector Machine and Multilayer Perceptron) required feature scaling to normalise the values of the input features using a common scale to increase convergence and performance.

Imbalance was present within the dataset between cancer-related and non-cancer-related classes, thus stratified training and testing partitions were employed in preserving the distribution of attributes between both partitions. This ensures that all models have trained and evaluated on a representative dataset and hence the performance estimates should be less biased.

C. Machine Learning Models

In this analysis, multiple machine learning methods were implemented to classify mutations found in the genome as either associated with cancer or not associated with cancer using information regarding single nucleotide polymorphism (SNP) feature data. Models that were utilized include: Logistic regression, Random forest, Weighted random forest, Support vector machine, Multilayer perceptron and XGBoost. These models were selected to provide an assessment of the performance of each algorithm utilizing a comparison based upon three different learning methods (linear, nonlinear, and ensemble).

Logistic regression is a simple and effective way of classifying data into two groups (binary classification), which is why it was selected as the baseline linear model. The Support vector machine classifier was used because it excels in classifying very large numbers of features (high-dimensional). The Multilayer Perceptron (a type of neural network) was selected because of its ability to model non-linear relationships between input features.

The Random forest and the Weighted random forest classifiers are both ensemble-based classifiers that build multiple trees (decision trees) to improve prediction accuracy and to reduce the overfitting of a model. The random forest model is especially beneficial when trying to model complex datasets, where the interactions of features are nonlinear. The XGBoost classifier is a type of boosting algorithm, which builds multiple classifiers sequentially, and therefore provides continual improvements to a model's accuracy over time by reducing classification errors.

All models were trained using the training dataset and evaluated using bootstrap resampling methods on the test dataset to evaluate model performance consistency.

D. Bootstrap Evaluation

Bootstrap resampling is a way to get accurate and consistent estimates of model quality by making many samples of the test data from the original sample with replacement (for instance, making an estimate of the mean horsepower for automobile models using only a small sample of automobiles from the original dataset).

One hundred bootstrap resampling iterations were conducted in this study. In each iteration, 30 samples of the test data were selected at random with replacement to use as the test set for the resampling process. The various models produced from the original training process were then used to generate predictions based on these samples of the test dataset.

Performance metrics (accuracy, precision, recall, F1 score, and specificity) were collected for each model across all iterations. The final reported performance metrics in this study are averages of the collected performance metrics over all bootstrap resampling iterations.

By using bootstrap resampling, the model evaluations become less dependent on one specific test data split and provide a more robust way of measuring how well the models perform on an independent dataset. Due to the nature of the bootstrap resampling process, there were minor differences in the metrics each iteration due to the stochastic nature of the bootstrap sampling process; however, the metric values remained consistent within a narrow range over the course of the bootstrap re-sampling iterations. This indicates that all of the models are stable.

E. Performance Metrics

The machine learning models used in this research were evaluated using various classification metrics so that an overall evaluation could be made of the effectiveness of the models in identifying cancer-associated mutations. Because the problem is one where mutations are classified either as being cancer-associated or as being non-cancer-associated, it is important to evaluate the overall correctness of the models as well as how well each class is identified. The classification metrics used in this evaluation include accuracy, precision, recall, F1 score and specificity.

- *Accuracy:* It represents the overall correctness of the model in classifying mutations. It indicates how many predictions were correctly made out of all predictions.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- *Precision: It* measures how many of the mutations predicted as cancer associated are actually cancer associated. It reflects the reliability of positive predictions.

$$Precision = \frac{TP}{(TP + FP)}$$

- *Recall:* measures how well the model identifies actual cancer associated mutations. It indicates the ability of the model to capture true positive cases.

$$Recall = \frac{TP}{(TP + FN)}$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

➢ *The F1 score:* provides a balance between precision and recall, especially useful when there is class imbalance. It combines both metrics into a single value.

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

➢ *Specificity:* measures how well the model correctly identifies non cancer associated mutations. It evaluates the model's ability to avoid false positives.

$$Specificity = \frac{TN}{(TN + FP)}$$

The four classification results (true positive, true negative, false positive and false negative) will describe a classification system's predictive abilities. A true positive will refer to a mutation that has been correctly classified as associated with cancer. A true negative will refer to a mutation that has been correctly classified as not being associated with cancer. A false positive will refer to a mutation that has been incorrectly classified as associated with cancer when, in fact, it is not. A false negative will refer to a mutation that has been incorrectly classified as not being associated with cancer when, in fact, it is. These values are used to calculate performance evaluation metrics, which can be retrieved from a confusion matrix.

F. Model Explainability (SHAP)

SHAP (SHapley Additive exPlanations) based explainability methods were utilized in order to improve the interpretability of machine learning models. SHAP is an important tool for allowing a clearer understanding of the way in which each feature affects the outcomes of a machine learning model's predictions or classifications (specifically, in this study, how the input features of a machine learning model are being used to determine whether or not a mutation has a better chance of being classified as being an associated with a cancer than does another mutation).

The SHAP analysis conducted in this study assessed the degree to which various features representing SNP feature scores (SIFT, PolyPhen2, Combined Annotation Dependent Depletion (CADD), and REVEL) were able to predict whether or not a particular mutation would be classified as associated with a cancer. The SHAP values denote whether a particular feature is contributing positively or negatively to whether a mutation would be predicted as associated with a cancer.

The analysis found that some features had more influence in contributing to the predictions of the models and, consequently, could be classified as being more important in relation to the determining of whether or not a mutation may have an impact. The higher the value associated with a given feature within the context of SHAP, the greater impact that feature is contributing toward making the classification decision, while the lower the value corresponding to that same feature, the lesser impact the feature will have in making the classification decision.

Through the implementation of SHAP based interpretability methodologies, the models are made to be more transparent and interpretable, therefore allowing individuals the ability to understand how decisions are made. Understanding how decisions are made, especially in the context of genomics, by determining what features are influencing decisions will help to provide biological insight into the meaning of mutations.

IV. RESULTS AND DISCUSSION

A. Performance Evaluation of Machine Learning Models

To ensure that the estimates of the performance of the machine-learning models that were implemented are accurate and stable, bootstrap resampling was used to evaluate the performance of the models, with the average of accuracy, precision, recall, F1 score and specificity calculated for each of the models over the multiple iterations used for bootstrapping.

Table 1 presents the comparative performance of all models based on the evaluation metrics. The ensemble models perform better than any other method. In addition, the Random Forest model produces the highest accuracy (approximately 83%) followed by XGBoost with an accuracy of approximately 82%. The Weighted Random Forest model also produces good results; however, it is very close in accuracy to the other two models.

Logistic Regression and other linear models failed to perform well, implying that these models lack the ability to handle complex non-linear relationships present in the feature data from SNPs. A similar situation happens with Multilayer Perceptron produced moderate performance, attributed to either the dataset size and/or model complexity.

Fig. 2 presents a comparative analysis of the performance of different machine learning models across evaluation metrics including accuracy, precision, recall, and F1 score. It can be observed that ensemble based models, particularly Random Forest and XGBoost, consistently achieve higher values across most metrics compared to other models. Random Forest shows the highest overall accuracy, while XGBoost performs closely with competitive results.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

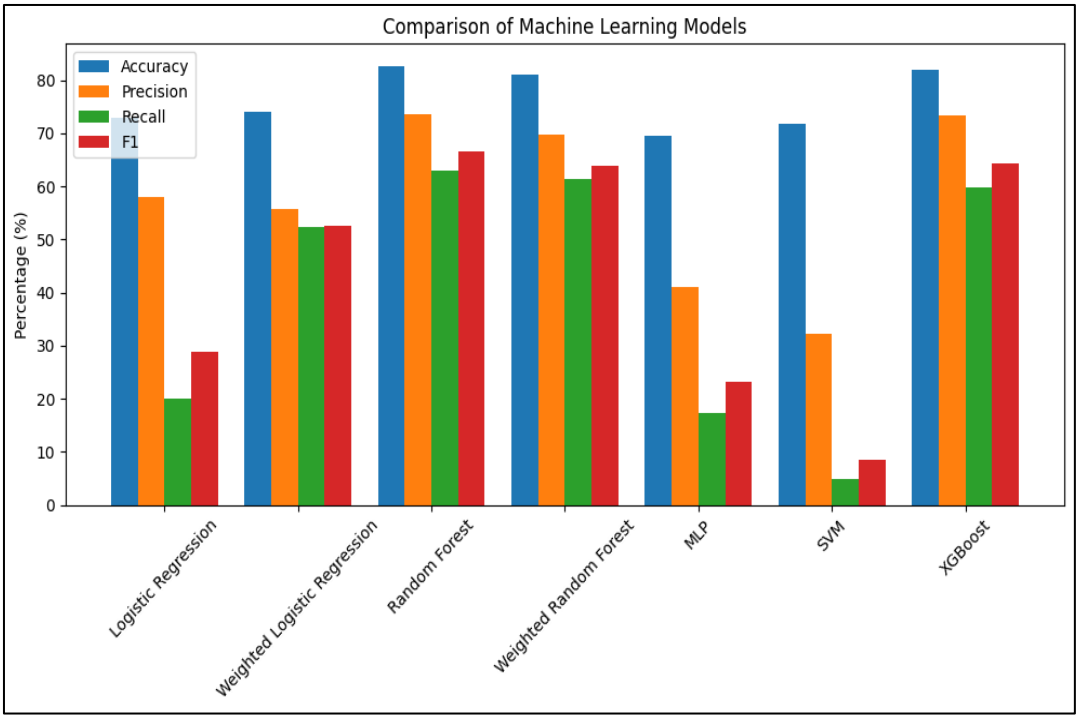

Fig. 2 Comparison of Machine Learning Models across Evaluation Metrics

In contrast, models such as Logistic Regression and Support Vector Machine demonstrate comparatively lower performance, indicating their limited ability to capture complex non-linear relationships present in SNP feature data. The Multilayer Perceptron model shows moderate performance across the metrics.

Additionally, it can be noted that recall values are relatively lower for some models, suggesting that certain cancer associated mutations are not correctly identified. This highlights the impact of class imbalance in the dataset and indicates the need for careful evaluation using multiple performance metrics.

Overall, Fig. 2 clearly demonstrates the effectiveness of ensemble learning methods in achieving better classification performance for cancer associated mutation prediction.

B. SHAP Based Explainability Analysis

In order to interpret the decision-making behavior of the models, SHAP-based analysis was performed. Using SHAP values, the contribution of each feature (i.e., SNPs) to the classification of mutations as cancer-associated or non-cancer-associated was quantified.

Each model had a set of SHAP summary values, summarized in Fig. 3, which provides a global (rather than per sample) summary perspective of the importance of each of the 14 features. It was noted during the SHAP analysis that the CADD and REVEL score features had the highest impact on the predictions of each of the different models and strongly determined whether or not a mutation was classified as cancer related.

Additionally, SHAP values were able to indicate the direction of impact made by each feature towards the likelihood of a mutation being classified as either cancer-associated or non-cancer-associated; that being the case, a feature with a positive SHAP value suggests that feature would increase the likelihood of being cancer-associated and a feature with a negative SHAP value suggests that that particular feature would contribute towards the mutation being classified as non-cancer-associated. This directional contribution provided an understanding of how the values of individual features influenced the way in which the models reached their respective decisions.

Table 1 Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	Specificity	F1
Logistic Regression	72.90	58.01	20.15	94.68	28.76
Weighted Logistic Regression	74.03	55.78	52.41	82.92	52.65
Random Forest	**83.3**	73.68	63.03	90.74	66.56
Weighted Random Forest	81.00	69.76	61.38	88.99	63.83
MLP	69.47	41.05	17.35	90.77	23.13
SVM	71.77	32.17	4.96	99.25	8.43
XGBoost	81.90	73.49	59.77	90.95	64.41

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

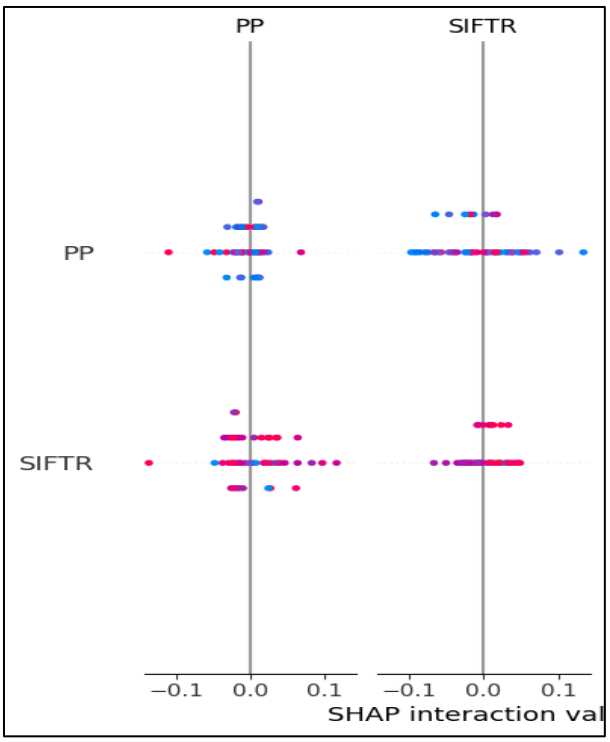

Fig. 3 SHAP Summary Plot Showing Feature Importance

V. CONCLUSION

This study presented a machine learning based approach for analyzing cancer associated mutations in the POLB gene using features derived from single nucleotide polymorphisms. Multiple classification models, including Logistic Regression, Random Forest, Support Vector Machine, Multilayer Perceptron, and XGBoost, were implemented and evaluated to identify the most effective method for mutation classification.

The experimental results demonstrated that ensemble based models, particularly Random Forest and XGBoost, achieved superior performance compared to other models, with accuracy values in the range of approximately 82 to 83 percent. The use of bootstrap resampling provided stable and reliable performance estimates, confirming the robustness of the models.

Furthermore, SHapley Additive exPlanations based analysis enhanced the interpretability of the models by identifying the most influential features contributing to classification decisions. This improves transparency and provides meaningful insights into the role of SNP features in cancer mutation prediction.

Overall, the findings indicate that machine learning techniques, combined with explainability methods, offer an effective approach for analyzing genomic data and predicting cancer associated mutations. The proposed methodology can support further research in bioinformatics and assist in improving the understanding of mutation driven diseases.

Future work can focus on improving model performance by incorporating larger and more diverse datasets, applying advanced deep learning techniques, and exploring additional feature engineering methods to enhance prediction accuracy.

REFERENCES

[1]. R. Alkhanbouli, A. Al-Aamri, M. Maalouf, K. Taha, A. Henschel, and D. Homouz, “Analysis of cancer-associated mutations of POLB using machine learning and bioinformatics,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 21, no. 5, pp. 1436–1444, 2024.
[2]. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[3]. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
[4]. P. C. Ng and S. Henikoff, “SIFT: Predicting amino acid changes that affect protein function,” Nucleic Acids Research, vol. 31, no. 13, pp. 3812–3814, 2003.
[5]. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, “A method and server for predicting damaging missense mutations,” Nature Methods, vol. 7, no. 4, pp. 248–249, 2010.
[6]. M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, and J. Shendure, “A general framework for estimating the relative pathogenicity of human genetic variants,” Nature Genetics, vol. 46, no. 3, pp. 310–315, 2014.
[7]. N. Ioannidis, V. J. Rothstein, V. Pejaver, J. Middha, S. McDonnell, J. Baheti, A. Musolf, H. Li, S. E. Pendergrass, D. A. Bick, et al., “REVEL: An ensemble method for predicting the pathogenicity of rare missense variants,” American Journal of Human Genetics, vol. 99, no. 4, pp. 877–885, 2016.
[8]. C. Cortes and V. Vapnik, “Support vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1516

[9]. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986

[10]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.