Quest Journals
Journal of Research in Pharmaceutical Science
Volume 8 ~ Issue 5 (2022) pp: 17-28
ISSN(Online) : 2347-2995
www.questjournals.org

Research Paper

Identification of the effect of missense variants associated with ribosomopathies using an ensemble approach to resolve conflicting interpretations

[1]Ajitha Mohan, [2]Sekar Kanagaraj
[1]*(Department of Computational and Data Sciences, Indian Institute of science, Bangalore, India)*
[2] *(Department of Computational and Data Sciences, Indian Institute of science, Bangalore, India)*
Corresponding Author: Ajitha Mohan

ABSTRACT: *The contemporary research studies suggest that the missense variants of ribosomes are the main cause of ribosomopathies. In specific, the conflict interpretation of the clinical significance of missense variants associated with ribosomopathies remains an unsolved issue. To solve such issue, the known information related to each missense variant like allelic score, selective sweep score, GERP score, grantham score, and pathogenicity predictive score is used to reclassify the variants tagged as 'Conflicting Interpretations'. Initially, the correlation between the aforementioned scores and the known effect of variants is identified. Based on the correlation study, the feature selection is done to generate various machine learning models such as support vector machine, logistic regression, KNN, random forest, XG Boost, and Decision Tree to identify the pathogenicity of variants. In this study, the random forest model is proposed as the best model based on performance metrics. The pathogenic variants predicted by the newly generated model can be used as a prognostic factor for further clinical study of ribosomopathies.*

KEYWORDS: *Missense variants, Machine learning Techniques, Ribosomopathy, Allele frequency, Clinical Significance.*

R*eceived 12 May, 2022;* ***R****evised 24 May, 2022;* ***A****ccepted 26 May, 2022 © The author(s) 2022.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

Due to the advent of next genome sequencing techniques, the numbers of genetic variants associated with ribosomopathies are very high. So there is a need of in silico approach to sort out the pathogenic variants from the prevailing list. Even though various pathogenicity prediction tools are proposed to analyze the variants, each tool has its own merits and drawbacks as they work on different properties. This is the point where the conflict of interpretation arises, as the different tool works on a unique property there is a difference in opinion. This problem is solved by the entry of the coincidence rule [1] this rule states that if the variant is identified as pathogenic by most of the available tools then that variant is nominated for a further clinical test. This rule is significantly added to the regulations of the Association for molecular pathology as well as the American college of medical genetics and genomics to interpret the causative variant for the diseased condition of the patients. To successfully apply this rule, integration techniques are applied whereby the decision of all the tools can be combined to a single score that represents the consolidated decision of all the available tools. As a recent trend machine learning techniques are used to practically solve the diagnosis of diseased conditions based on appropriate feature selection [2]. Likewise, such ensemble techniques are employed to train the new model to calculate the integrated score, thereby the deleterious variant can be identified [3]. As of now various integrated tools were launched successfully such as metaSVM, CONDEL, CoVEC, IMHOTEP [4] which are available as webservers for small query and in executable form for large queries which runs locally.

Mostly, pathogenicity prediction tools work on the concept of sequence conservation information at specific functional sites. One such successful tool is SNP Effect Predictor which is launched by the ENSEMBLE project [5] to access the conservative information from all the species instantly [6]. Likewise, CONDEL predictors calculate their integrated score using four different methods such as SVS, SAS, WVS, WAS, and finally concluded that WAS method shows significant results compared to that of other available predictive scores [7]. Similarly CAROL calculates the integrated score by combining the probabilistic score of SIFT and POLYPHEN and proves that ensemble method shows the better performance than the individual method [8]. Subsequently, a new model is developed using Support Vector Machine and Linear Regression method with linear, radial and polynomial kernel and finally proved that ensemble methods are better predictors than the individual predictor [9]. Besides, the Mutation Assessor-2 tool is capable of processing the VCF files automatically and they apply ensemble Bayes classifiers to predict the pathogenicity of the query variant [10]. In the same way, CADD includes the ensemble notations of several genomic features to calculate the pathogenicity score for all the single nucleotide polymorphism found in the reference assembly and this score is proved to be highly correlated with Mendelian diseases [11]. As a successful implementation in 2017, CADD score is employed to pick up 11 patients from 238 breast and ovarian cancer patients for further clinical studies. In addition to the CADD score, the population frequency range also plays an important role in picking up the susceptible variants of cancer patients [12]. FATHMM is capable of calculating the pathogenicity score for snp derived from non-human and human sequencing projects [13]. As an update, FATHMM-V.2.3 is designed based on a weight scoring scheme in such a way it doesn't require prior information regarding the protein function [14]. PredictSNP released in 2014 is designed to predict deleterious snp from the curated training dataset by excluding the duplicate and other discrepancy data sets, thereby it fills the gap between the benchmark data sets and training data sets [15]. In 2020, the study of family-specific variants reveals that the sensitivity of particular variant is correlated with their occurrence in a unique protein domain [16]. Additionally, the tumorogenesis role of snp (rs1800371 and rs1042522) on p53 protein encoded by TP53 gene is examined and confirmed as pathogenic by the ensemble approach [17]. Even though different algorithm works with different score, the phylogenetic and conservative scores are proved to play a major role in pathogenicity prediction [18]. Finally, this study highly recommends the combination analysis of scores from different pathogenicity predictor tools in Healthcare Applications. It is noteworthy that certain information like allele frequency (Minor Allele Frequency), GERP (Genomic Evolutionary Rate Profilling), and grantham score also play an important role in deciding the pathogenicity of variants [19]. The susceptible variants associated with specific disease can be identified by prioritizing the scores predicted by insilico pathogenic predictors [20]. In particular, NSS score stands for the negative selective sweep score that tells us the information about the presence and absence of ancestral alleles of a variant in a particular gene. As negative sweep score is closely linked to pathogenicity and disease like schizophrenia, they are prioritized in the study of genetic disorders [21]. All the above research works show the power of ensemble techniques and their success stories in the medical field. To this end, all the aforementioned scores are utilized as a key factor in the present study to resolve the conflicting interpretations associated with ribosomopathies.

II. MATERIALS AND METHODS

The main objective of the present study is to reclassify the effect of variants tagged as 'conflicting interpretations' by utilizing the known information associated with benign and pathogenic variants of ribosomopathies. As 80s ribosomes are mostly associated with ribosomopathies, the corresponding 8703 genes were retrieved from the GeneCards database which acts as a repository for all the data related to human genes [22]. The retrieved lists of genes were submitted to the Clinvar database, which contains information related to the clinical significance of all variants of a specific gene [23]. Thus the lists of missense variants of each gene and their corresponding clinical significance were retrieved. Among all the categories the missense variants were alone chosen as it plays a vital role in NGS data more than 60%. The missense variants with benign and pathogenic effects were filtered out from other variants and the complete workflow is shown in Figure. 1. As the number of genes involved in ribosomopathies are high, the data collection and curation are done by writing the python scripts. Eventually, after removing the redundant data, 6081 benign and 7960 pathogenic variants were alone sorted along with the details of the chromosomal position and their corresponding rs-ID. The corresponding scores of each variant such as minor allele frequencies, GERP, grantham score, NSS score, and pathogenicity scores are retrieved using python scripts. The threshold of each score is used as a key criterion to interpret the effect of variants. Accordingly, such threshold of each score differs as they are related to different attributes of variant. In the case of NSS score, their negative values are substantiated to be closely related to disease-causing variants. Such selective sweep scores of the variants were sorted out from the UCSC-data integration pool [24]. Post that, to filter out the negative sweep score NSSscore script is written in python and deposited in the github repository for ease of use. It is capable of analyzing more than 1000 data sets and thus it breaks the limitation of the number of query data and its final results are represented in graphical format. Similarly, variants with less than 0.1% minor allele frequency are considered as significant as they are deployed in the study of linkage disequilibrium in most cases. The variants with higher grantham scores are targeted as they are reported as non-tolerable substitutions. Similarly, as the Genomic Evolutionary Rate Profiling score indicates the level of conservation in each species, the variants with a high GERP score are specifically targeted. The pathogenicity scores predicted by various tools such as SIFT[25], POLYPHEN[26], CADD, REVEL[27], METALR, METAL ASSESSOR[28] are included as they are highly recommended by the ensemble genome project. Further, all the above statistical scores related to each variant were retrieved and analyzed using python script. All the above information was used for a correlation study between the effects of variants and their corresponding statistical scores. To perform the correlation study, initially, the effects of variants are converted to numerical binary output. Then the correlation between the clinical significance of the variant and their statistical scores was studied by identifying Pearson's correlation coefficient value using Equation. (1).

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}} \qquad \rightarrow \qquad (1)$$

Based on the correlation studies, scores that are positively correlated with the effect of variants are identified. Remarkably, such scores are utilized as input datasets to generate various machine learning models using logistic regression, support vector machine, KNN, XG Boost, Decision Tree, and random forest. In each approach, a standard machine learning pipeline is followed to solve the binary classification problem and it is successfully executed using python as the development environment. It is noteworthy that the pathogenic scores and the effect of variants are considered as attributes and label sets respectively in the present study. As the range of the score differs, initially standardscalar class is applied to perform the feature scaling task before training the machine learning models. Further, the data is divided into training and test data sets in 80% and 20% ratios respectively. To evaluate each model various metrics such as precision (Equation. (2)), recall (Equation. (3)), F1-Score (Equation. (4)), MCC value (Equation. (5)) and their corresponding accuracy(Equation. (6)) are calculated. Finally, ROC and AUC curves were also plotted to identify the best model.

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad \rightarrow \qquad (2)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \qquad \rightarrow \qquad (3)$$

$$\text{F1 Score} = \frac{2*(\text{Recall}*\text{Precision})}{(\text{Recall}+\text{Precision})} \qquad \rightarrow \qquad (4)$$

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad \rightarrow \qquad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \qquad \rightarrow \qquad (6)$$

III. RESULTS AND DISCUSSION

The Pearson's coefficient value between the effect of the variant and its corresponding scores such as minor allele frequency, GERP, grantham, NSS, and pathogenicity score for all the missense variants have been calculated and analyzed to find the strength of association between the variables. However, NSS score data is not available for most of such variants and therefore no correlation is identified. But NSS scores for available benign and pathogenic variants associated with ribosomopathies are plotted and shown in Figure. 2. Likewise among all the scores, the pathogenicity score is identified to have a strong positive correlation with the effect of the variant as it has a higher Pearson's correlation coefficient (r=0.80) whereas the remaining scores show a weak correlation (r<0.5) as shown in Figure. 3. Eventually, based on the pathogenicity scores, new statistical models are generated using various machine learning approaches such as KNN, XG Boost, Decision Tree, support vector machine, and random forest. To sort out the best model, all the statistical metrics of each model are calculated and shown in Figure. 4.

Based on the combination of accuracy and ROC/AUC curve (Figure. 5), the random forest model is identified as the best model as it has consolidated higher performance metrics compared to other models. Even though the bagging or bootstrap aggregation method is implemented in the random forest model, tunning of hyperparameters is done additionally to avoid the overfitting and underfitting problems. During such optimization process, various hyperparameters like criterion, depth of the tree, number of samples to split internal node, number of samples at a leaf node, a weighted fraction at a leaf node, number of features, and number of leaf nodes are fine tunned by applying the grid search algorithm. Especially “Gini” criterion is applied to measure the Gini impurity by using Equation. (7) as it is more vivid than the entropy measurements (Equation. (8)) in terms of computational complexity. In this way, the Gini impurity is confirmed to be zero at the leaf node and subsequently, the same process is iterated to all decision trees to navigate the fine tunned classification algorithm. The remaining optimum hyperparameters are also derived by executing the grid search algorithm and in addition cross-checking is also done by identifying the optimized value of the individual hyperparameter at which the tree overfits the training data as shown in Figure. 6. However, to avoid the overfitting and underfitting problems, the accuracy of both the training set and the test sets are predicted and their difference is calculated as 1%.

$$I_G = 1 - \sum_{j=1}^{C} p_j^2 \quad \rightarrow \quad (7)$$

$$I_H = -\sum_{j=1}^{c} p_j \log_2(pj) \quad \rightarrow \quad (8)$$

Furthermore, to validate the random forest model, the K-fold cross-validation method is adopted. Consequently, the input data is split into 10 folds (K=10), such that we have 10 different sets of training and test data to build the classification model. To assess the efficiency of the model, the combination of training and test data set differs in each iteration thereby the usage of all the data is confirmed. Such a combination confirms that the model eradicates the underfitting and overfitting problems with low bias and low variance. Further, as the experiment is iterated with 10 different holdout sets, ten different accuracies are calculated as follows 0.96352313, 0.96441281, 0.96081923, 0.97506679, 0.95636687, 0.96438112, 0.97061443 0.96794301, 0.96794301, and 0.96616207. Finally, the mean of all accuracy is calculated as 96.5% and thus the model is fine-tuned to classify the unknown data. The following Figure. 7 represents the confusion matrix of the random forest model.

3.1 Ribosomopathy Variants

The list of 58 Genes related to ribosomopathies was retrieved from gene cards and their corresponding variants tagged as ‘conflicting interpretations’ were retrieved from the clinvar database. In this study, a newly generated random forest model is deployed to resolve the conflicting interpretations. Among all the genes, TP53 contains more number of variants (139) tagged as ‘conflicting interpretations’ where 39% and 61% of variants are reclassified as benign and pathogenic respectively using a novel random forest model. Eventually, the variants of TP53 gene that are identified as pathogenic can be further used as a target in drug designing process as it have major role in ribosomopathies and cancer. Furthermore, in RPL5, DDX41, POLR1A genes 100% of variants tagged as ‘conflicting interpretations’ are reclassified as pathogenic. The predicted classes of remaining genes are shown in the supplementary material.

IV. CONCLUSION

To reclassify the ribosomopathy variants tagged as ‘conflicting interpretations’, various machine learning models are generated, and finally based on the performance metrics random forest model is identified as the best model. As this study includes different information of variants like computational and predictive data, allelic data, and population data, their individual drawbacks are eradicated by the ensemble approach. This study aids clinical genomist to resolve the conflict interpretations raised by the clinvar submitters in predicting their clinical significance. Furthermore, the present work helps to overcome the caveats of study of the polygenic disease. Also, this study acts as supportive evidence in fixing the genomic medicine, thereby it avails the clinicians to identify the exact drug target that ultimately leads to the effective treatment for ribosomopathies.

DATA AVAILABILITY

The data(supplementary) we generated is available at github repository (https://github.com/project6656/ribo) , which can be used with proper citation.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no competing interest that influence the proposed work reported in this paper.

ACKNOWLEDGEMENTS

The author (AM) thanks Dr. D.S. Kothari Postdoctoral Fellowship (DSKPDF), funded by the University Grants Commission (UGC). The authors thank Department of Computational and Data Sciences, Indian Institute of science, Bangalore, India for providing necessary support.

REFERENCES

[1]. De la Campa, E. Á., Padilla, N., & de la Cruz, X, Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence. BMC genomics, (2017). **18**(5), 1-14.

[2]. Muhammed Niyas, K. P., & Thiyagarajan, P., Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification. Journal of King Saud University-Computer and Information Sciences, 2021.

[3]. Lai, J., Yang, J., Gamsiz Uzun, E. D., Rubenstein, B. M., & Sarkar, I. N, LYRUS: a machine learning model for predicting the pathogenicity of missense variants. Bioinformatics advances, 2022. **2**(1), vbab045.

[4]. Knecht, C., Mort, M., Junge, O., Cooper, D. N., Krawczak, M., & Caliebe, A., IMHOTEP—a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. Nucleic acids research, 2017. **45**(3), e13-e13.

[5]. Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., & Flicek, P, Ensemble. Nucleic acids research, 2018. **46**(D1), D754-D761.

[6]. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F, Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 2010. **26**(16), 2069-2070.

[7]. Gonzalez-Perez, A., & López-Bigas, , Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. The American Journal of Human Genetics, 2011. **88**(4), 440-449.

[8]. Lopes, M. C., Joyce, C., Ritchie, G. R., John, S. L., Cunningham, F., Asimit, J., & Zeggini, E. A combined functional annotation score for non-synonymous variants. Human heredity, 2012. **73**(1), 47-51.

[9]. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human molecular genetics, 2015. **24**(8), 2125-2137.

[10]. Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. Nature methods, 2014. **11**(4), 361-362.

[11]. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, 2019. **47**(D1), D886-D894.

[12]. Nakagomi, H., Mochizuki, H., Inoue, M., Hirotsu, Y., Amemiya, K., Sakamoto, I., & Omata, M. Combined annotation-dependent depletion score for BRCA1/2 variants in patients with breast and/or ovarian cancer. Cancer science, 2018. **109**(2), 453-461.

[13]. Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., & Gaunt, T. R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human mutation, 2013. **34**(1), 57-65.

[14]. Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N., & Gaunt, T. R. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Human genomics, 2014. **8**(1), 1-6.

[15]. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., & Damborsky, J.. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. 2014, PLoS computational biology, **10**(1), e1003440.

[16]. Zaucha, J., Heinzinger, M., Tarnovskaya, S., Rost, B., & Frishman, D. Family-specific analysis of variant pathogenicity prediction tools. 2020. NAR Genomics and Bioinformatics, **2**(2), lqaa014.

[17]. Bó, I. D., Teodoro, L., Peres, K. C., Teixeira, E. S., Bufalo, N. E., & Ward, L. S. MON-519 In Silico Analysis of rs1042522 and rs1042522 Polymorphic Variants of TP53 Gene. Journal of the Endocrine Society, 4(Supplement_1), 2020, MON-519.

[18]. Sun, H., & Yu, G.. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. Scientific reports, 2019. **9**(1), 1-11.

[19]. Ganakammal, S. R., & Alexov, E. Evaluation of performance of leading algorithms for variant pathogenicity predictions and designing a combinatory predictor method: application to Rett syndrome variants. PeerJ, 2019. 7, e8106.

[20]. Ganesh, S., Ahmed P, H., Nadella, R. K., More, R. P., Seshadri, M., Viswanath, B., & Rao, M.. Exome sequencing in families with severe mental illness identifies novel and rare variants in genes implicated in Mendelian neuropsychiatric syndromes. Psychiatry and clinical neurosciences, 2019. **73**(1), 11-19.

[21]. Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A. J., & International Headache Genetics Consortium.. Genetic markers of human evolution are enriched in schizophrenia. Biological psychiatry, 2016. **80**(4), 284-292.

[22]. Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., & Lancet, D.. The GeneCards Suite. In Practical Guide to Life Science Databases, Springer, Singapore. 2021, (pp. 27-56).

[23]. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R.. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, 2014. **42**(D1), D980-D985.

[24]. Hinrichs, A. S., Raney, B. J., Speir, M. L., Rhead, B., Casper, J., Karolchik, D., & Kent, W. J. UCSC data integrator and variant annotation integrator. Bioinformatics, 2016, **32**(9), 1430-1432.

[25]. Ng, P. C., & Henikoff, S.. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 2003, **31**(13), 3812-3814.

[26]. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., & Sunyaev, S. R., A method and server for predicting damaging missense mutations. Nature methods, 2010, **7**(4), 248-249.

[27]. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., Powell, I. J. & Sieh, W. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American journal of human genetics, 2016, **99**(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

[28]. Reva, B., Antipin, Y., & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research, 2011. **39**(17), e118-e118.

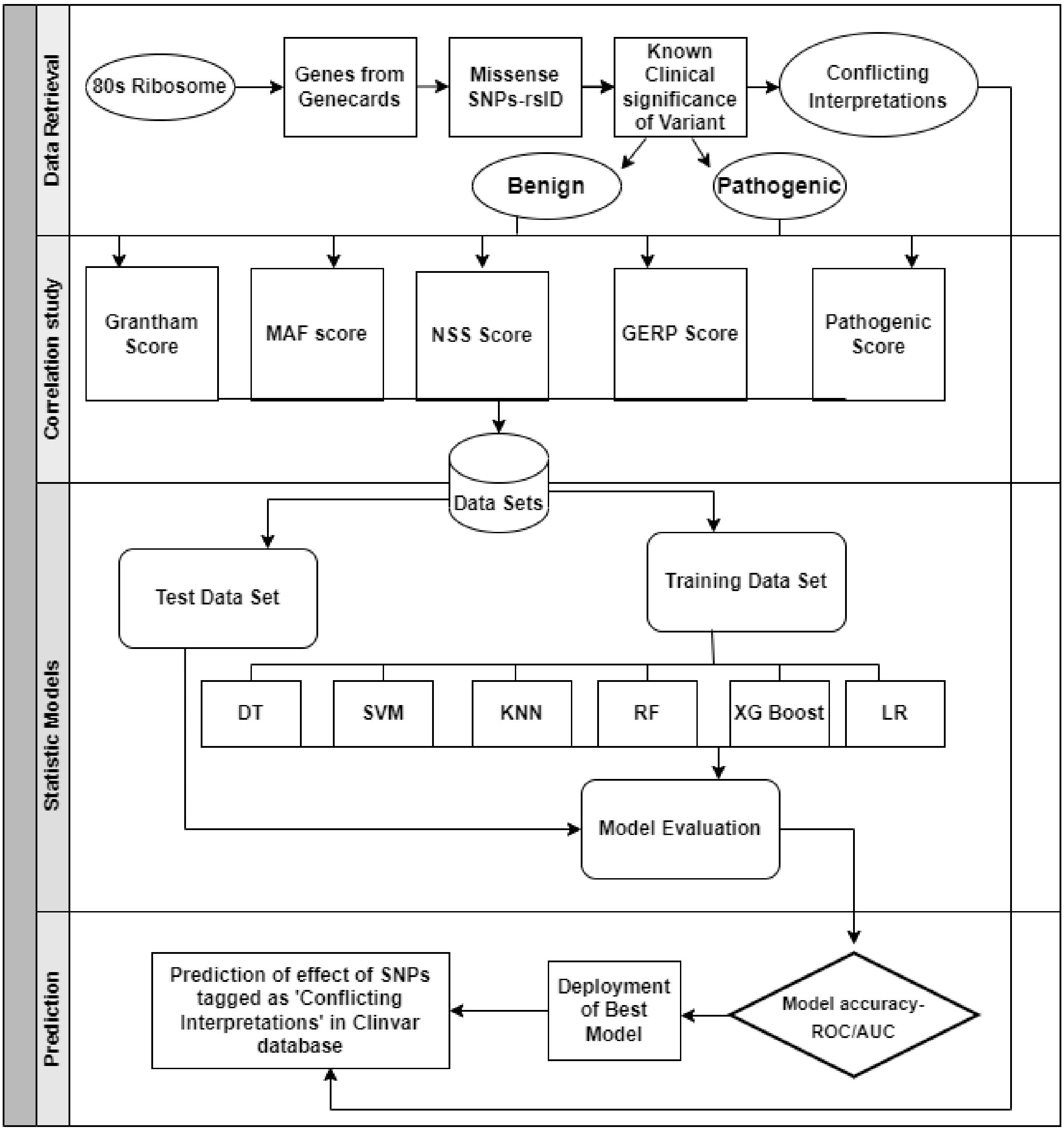


Figure. 1 Representation of the work flow.

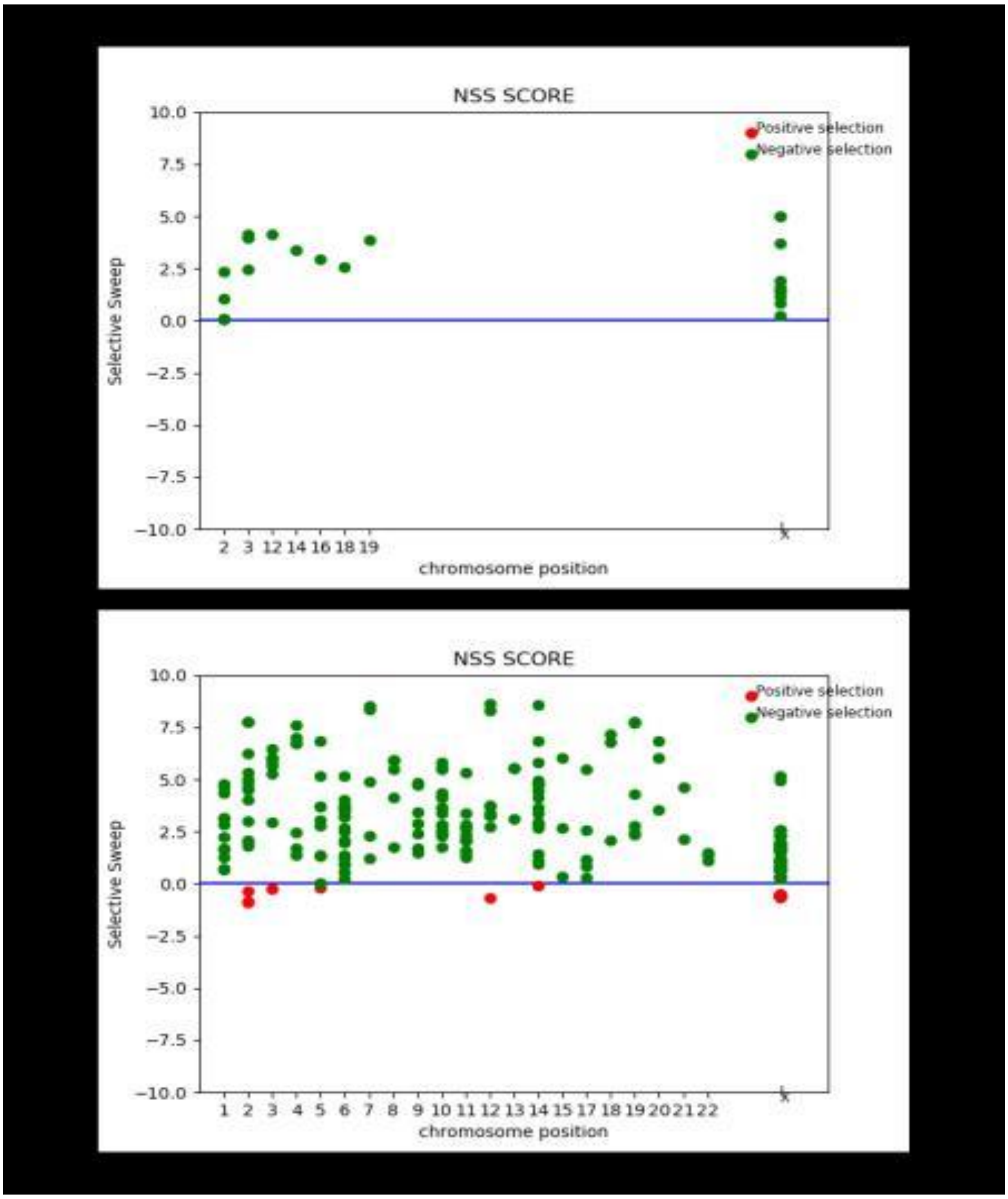


Figure. 2 a) Representation of NSS Score for Pathogenic variants b) Representation of NSS Score for Benign variants.

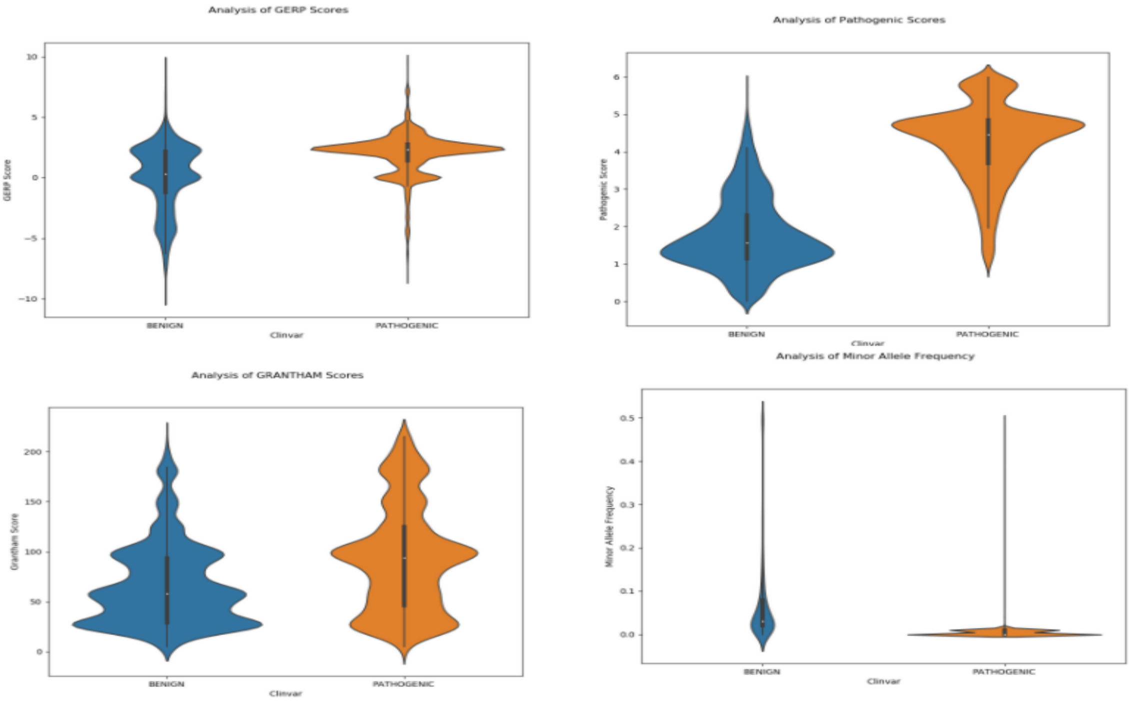
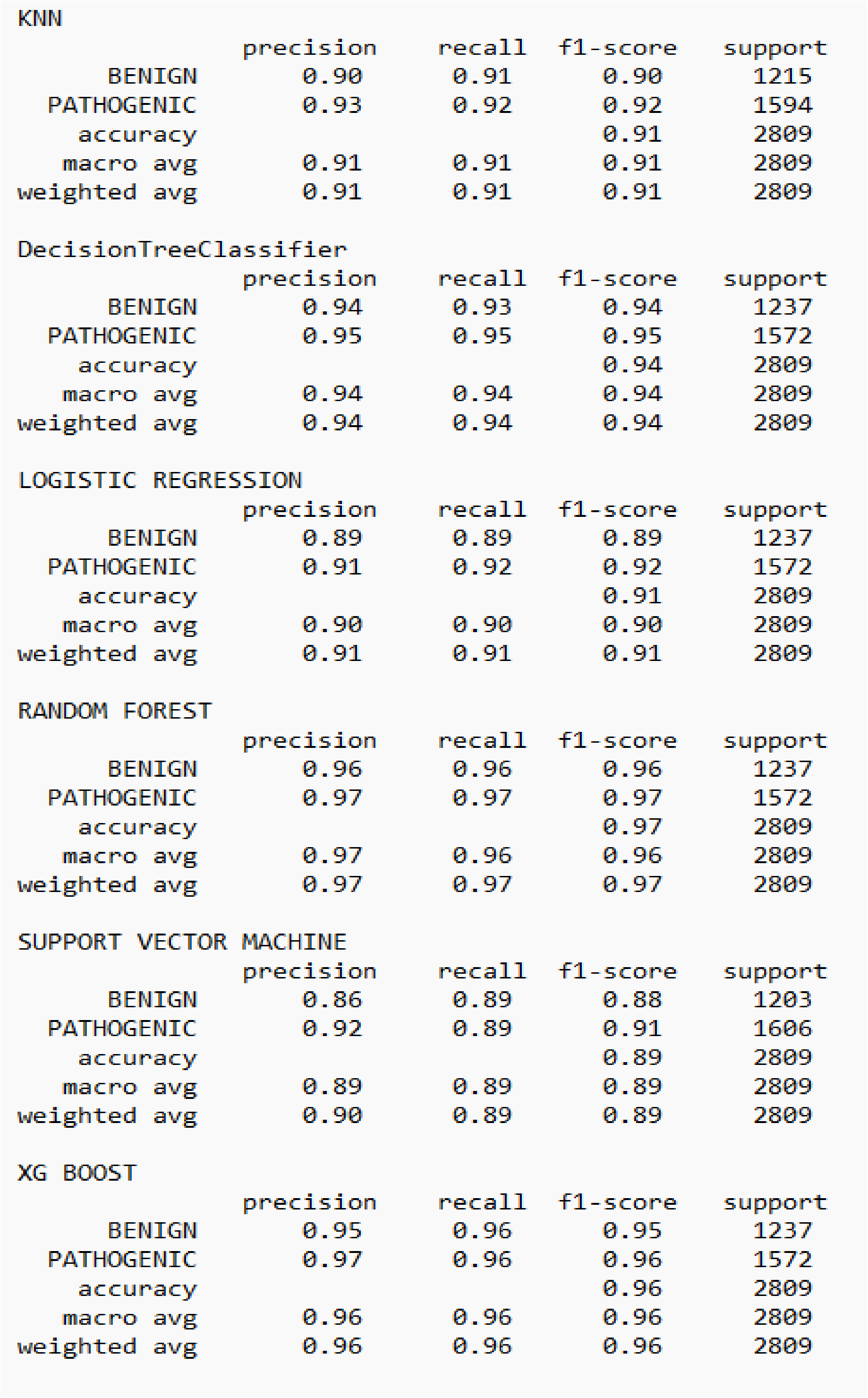


Figure. 3 Violin plot of the Pearson's Correlation between the effect of variants and statistical scores.



```
KNN
              precision    recall  f1-score   support
      BENIGN       0.90      0.91      0.90      1215
  PATHOGENIC       0.93      0.92      0.92      1594
    accuracy                           0.91      2809
   macro avg       0.91      0.91      0.91      2809
weighted avg       0.91      0.91      0.91      2809

DecisionTreeClassifier
              precision    recall  f1-score   support
      BENIGN       0.94      0.93      0.94      1237
  PATHOGENIC       0.95      0.95      0.95      1572
    accuracy                           0.94      2809
   macro avg       0.94      0.94      0.94      2809
weighted avg       0.94      0.94      0.94      2809

LOGISTIC REGRESSION
              precision    recall  f1-score   support
      BENIGN       0.89      0.89      0.89      1237
  PATHOGENIC       0.91      0.92      0.92      1572
    accuracy                           0.91      2809
   macro avg       0.90      0.90      0.90      2809
weighted avg       0.91      0.91      0.91      2809

RANDOM FOREST
              precision    recall  f1-score   support
      BENIGN       0.96      0.96      0.96      1237
  PATHOGENIC       0.97      0.97      0.97      1572
    accuracy                           0.97      2809
   macro avg       0.97      0.96      0.96      2809
weighted avg       0.97      0.97      0.97      2809

SUPPORT VECTOR MACHINE
              precision    recall  f1-score   support
      BENIGN       0.86      0.89      0.88      1203
  PATHOGENIC       0.92      0.89      0.91      1606
    accuracy                           0.89      2809
   macro avg       0.89      0.89      0.89      2809
weighted avg       0.90      0.89      0.89      2809

XG BOOST
              precision    recall  f1-score   support
      BENIGN       0.95      0.96      0.95      1237
  PATHOGENIC       0.97      0.96      0.96      1572
    accuracy                           0.96      2809
   macro avg       0.96      0.96      0.96      2809
weighted avg       0.96      0.96      0.96      2809
```

Figure. 4 Screenshot of the performance metrics calculated by machine learning models.

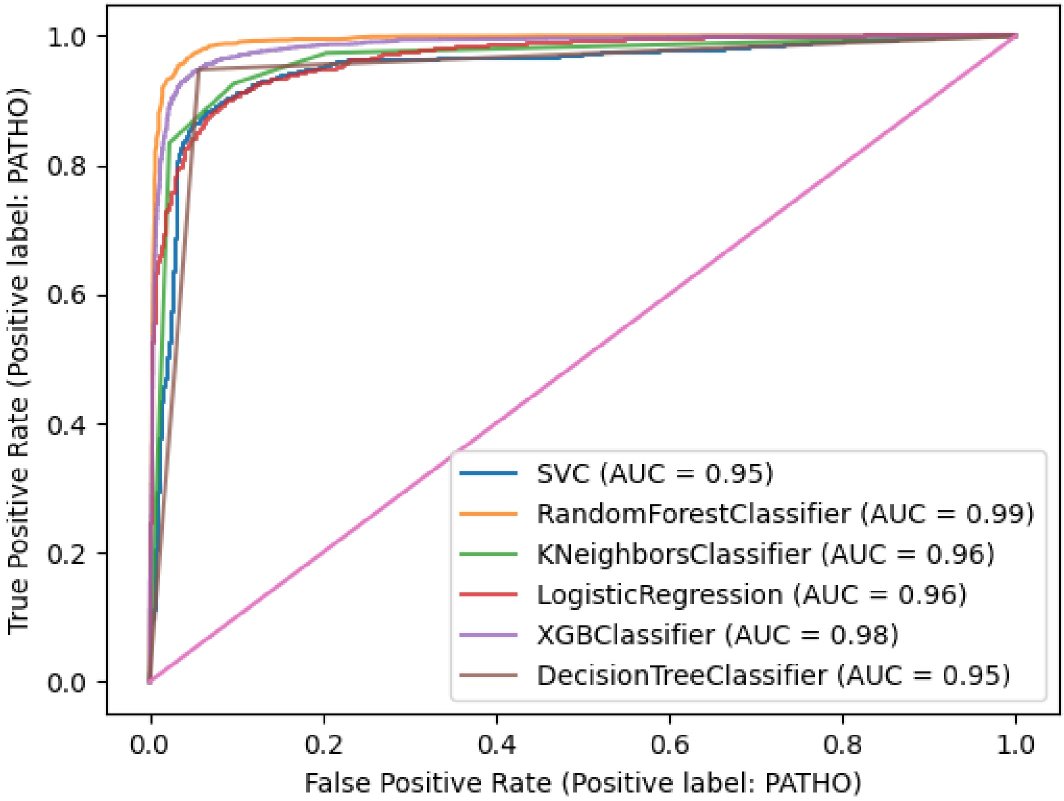


Figure. 5 Representation of Receiver Operating Characteristics.

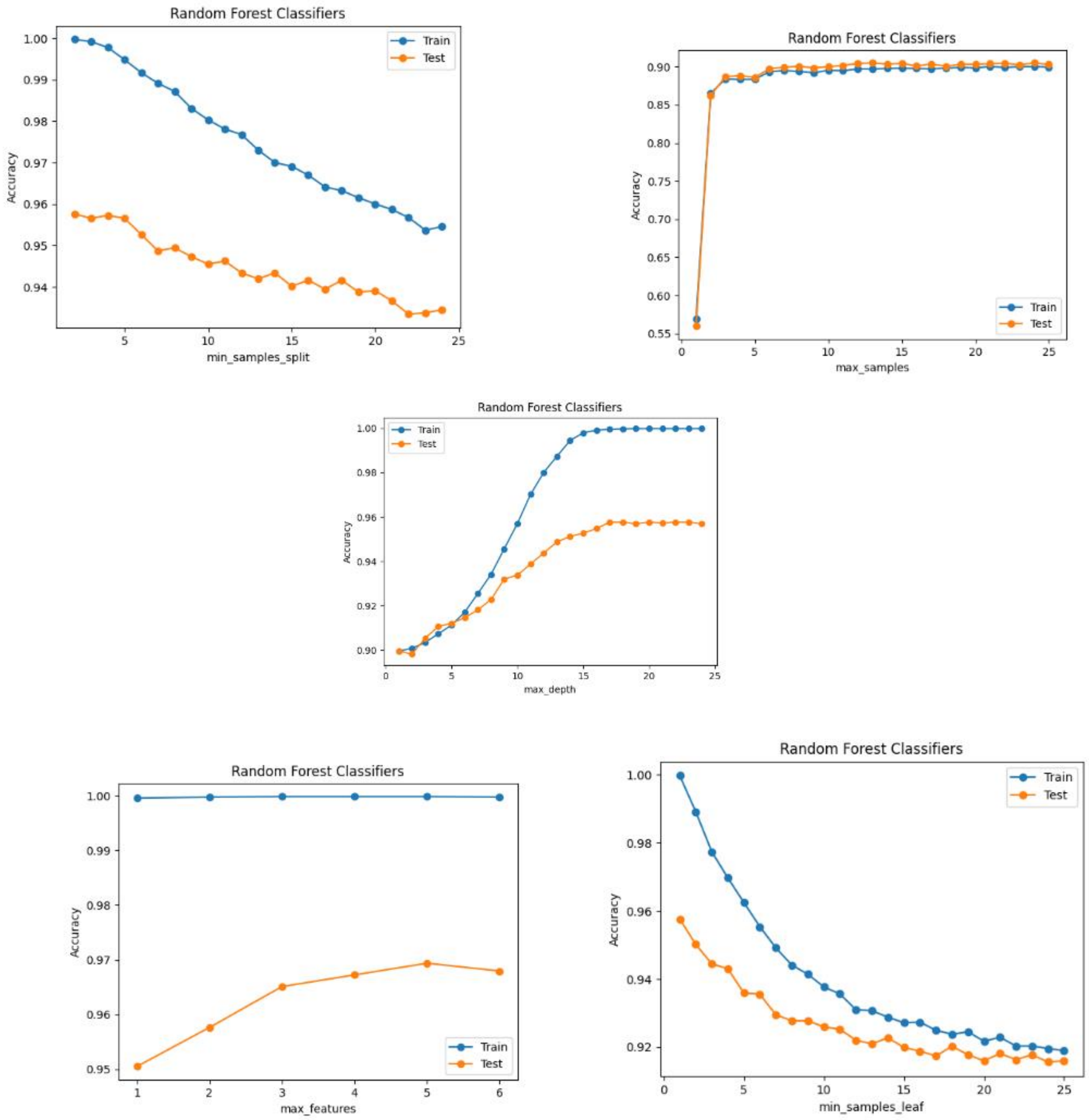


Figure. 6 Hyperparameter tuning for Random Forest model.

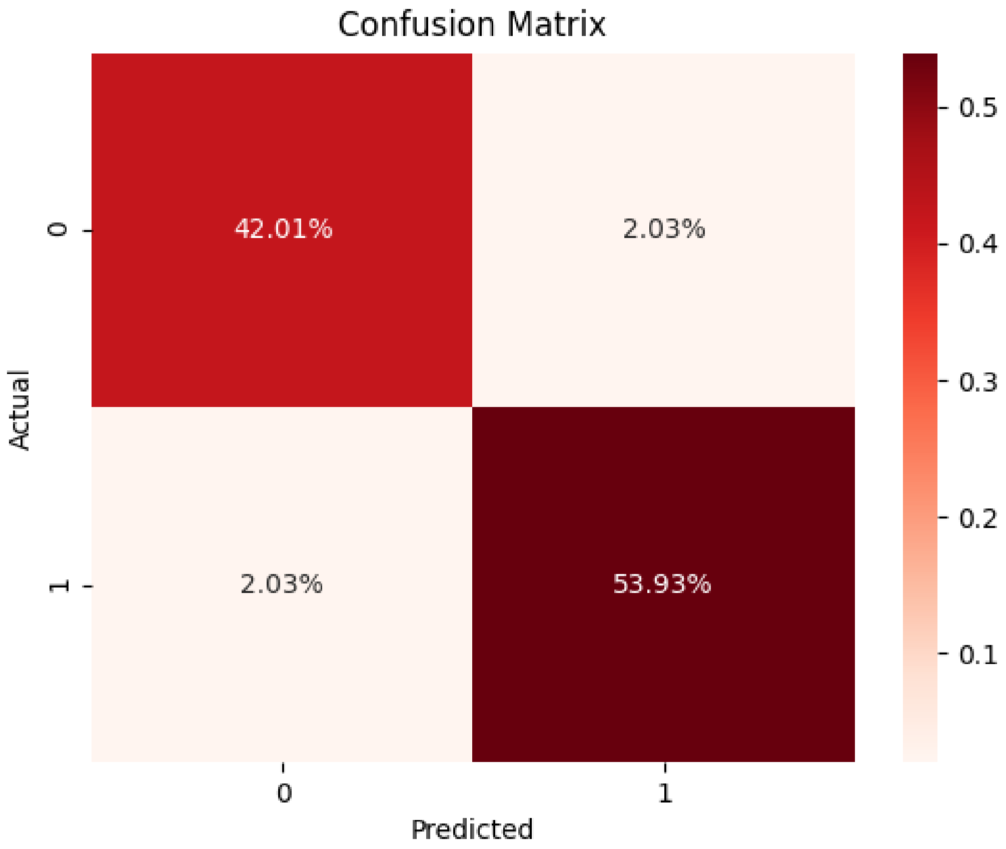


Figure. 7 Representation of the confusion matrix