International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparison of Various Variant Prediction Algorithms and Conservation Scores by Machine Learning Approaches

Sombuddha Roy Bhowmick

Abstract:
Machine learning methods are widely used for prediction of the pathogenicity and deleteriousness of the different types of single nucleotide variants (SNVs). This study encompasses a preliminary approach towards the application of different types of machine learning algorithms on the dbNSFPv4.7 (database of Nonsynonymous SNPs Functional Predictions) database to compare the various prediction algorithms and conservation scores to determine which one performs the best. This paves the way for selecting the best prediction algorithm or the combination of the better prediction algorithms in annotation of the variants and classifying them as pathogenic or likely pathogenic, benign or likely benign. The different learning approaches were compared based on the metrices like, accuracy, recall, precision and F1 score. Among the different variant prediction algorithms and conservation scores available in the dbNSFP database, BayesDel AF was the best perfoming followed by BayesDel noAF, MetaRNN, MCAP and MutPred. bStatistic, fitCons, fathmm XF, MutationTaster and phastCons17way primate performed the worst among 31 prediction algorithms and 9 conservation scores selected for this study. The best performing features or instances can be prioritized over others while selecting the functional prediction algorithms, which are used to determine the driver mutations or the pathogenic variants from omics datasets.

Keywords: Machine Learning, Prediction Algorithms, Conservation Scores, Single Nucleotide Variants (SNVs), dbNSFP

Background:
Machine learning algorithms employ a variety of methods like – statistical, probabilistic and optimizations to learn from the data and detect patterns from large, complex and unstructured datasets. There are four types of machine learning algorithms: unsupervised, semi-supervised, supervised and reinforcement. These algorithms are widely used and supervised learning methods are mainly used for disease modelling and variant effect prediction. In supervised learning, a prediction model is first developed by learning from a pre-existing dataset, which is then applied to predict the outcomes of unlabeled targets in the datasets. Neural networks, sometimes called ANNs (Artificial Neural Networks) are a subset of machine learning, and they are at the heart of deep learning models. They rely on the training data and improves the accuracy over time.
The scope of this paper is primarily based on the comparison of different variant prediction algorithms and conservation scores to predict the pathogenicity and deleteriousness of a variant. The scores predicted by the different prediction algorithms and their ability to correctly predict the clinical impact of those variants are assessed in this research. For this purpose, data from the dbNSFP database has been used.

dbNSFP is a database which has been developed for the functional prediction and annotation of all the potential non-synonymous single nucleotide variants (nsSNVs) and splice site single nucleotide variants (ssSNVs) of the human genome. dbNSFPv4.7 has a total of 84013490 nsSNVs and ssSNVs. It compiles prediction scores from 43 different prediction algorithms like: SIFT, SIFT4G, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaLR, MetaSVM, MetaRNN, CADD, CADD_hg19, VEST4, PROVEAN, FATHMM-MKL coding, FATHMM-XF coding, fitCons x 4, LINSIGHT, DANN, GenoCanyon, Eigen, Eigen-PC, M-CAP, REVEL, MutPred, MVP, gMVP, MPC, PrimateAI, GEOGEN2, BayesDel_AF, BayesDel_noAF, ClinPred, LIST-S2, VARITY, ESM1b, EVE, AlphaMissense, ALoFT, 9 conservation scores: PhyloP x 3, phastCons x 3, GERP++, SiPhy and bStatistic) and other related information including allele frequencies of various cohorts and population, functional description of genes, gene interaction information and others.

Materials & Methods:

The dbNSFPv4.7 which is based on the Gencode release 29 / Ensembl version 94 was downloaded and the information of 31 prediction algorithms and 9 conservation scores were extracted from the database. The clinical implications of these variants on human health are also present in this database, which has been collected from the ClinVar database. The level of confidence and the classification of the variants are largely dependent on the supporting evidence of the variants in the ClinVar database.

The prediction algorithms and their raw scores included in this study: SIFT, SIFT4G, AlphaMissense, BayesDel_AF, BayesDel_noAF, DANN, deogen2, Eigen_raw, Eigen_PC, ESM1b, EVE, FATHMM, FATHMM_MKL, FATHMM_XF, fitCons, gMVP, LIST_S2, LRT, MCAP, MetaLR, MetaRNN, MetaSVM, MPC, MutPred, MVP, MutationAssessor, MutationTaster, PrimateAI, PROVEAN, VARITY_ER and VARITY_R. The 9 conservation scores considered in this study: bStatistic, GERP++, SiPhy, phastCons100way_vertebrate, phastCons17way_primate, phastCons470way_mammalian, phyloP100way_vertebrate, phyloP17way_primate and phyloP470way_mammalian.

The variants missing any ClinVar values were excluded from the study, as this would incorporate inaccuracy within the models. For comparison of the different machine learning algorithms, the large dataset was divided into multiple smaller datasets comprising of a particular prediction algorithm scores or conservation scores and their corresponding ClinVar values.

The values in ClinVar were encoded into four separate classes – LB (Likely_Benign), B (Benign), LP (Likely_Pathogenic) and P (Pathogenic). For ClinVar annotations, where more than one clinical annotation was attached, the first value was referred and it was classified among the four encoded classes. The rest of the annotations of ClinVar like “Drug_Response”, “Risk_Factor” or “Confers_Sensitivity” and others, which did not have any of the four encoded classes were not considered in this study.

In each of the smaller datasets, the variants missing prediction algorithm scores or conservation scores were removed.

The supervised machine learning approaches as well as neural network models were used to train the dataset. Among the supervised learning algorithms, Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Gaussian Naïve Bayes, Extreme Gradient Boosting (XGBoost), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) were used. MLP (Multi-Layer Perceptron) classifier, which is a feedforward Artificial Neural Network (ANN) model was also used for training the dataset. The dataset was divided into 70:30 ratio of test and training subset. Data standardization procedures like scaling, transformation were performed on each of the

datasets, on which the algorithms were to build a predictive model. The target labels i.e. ClinVar values were encoded as integers as some of the algorithms like QDA and LDA require so.

Results:

Accuracy: It is a metric to determine how often a machine learning model correctly predicts the outcome. Accuracy can be calculated by dividing the number of correct predictions (true positives and true negatives) by the total number of predictions. It is a general metric that considers both positive and negative predictions.

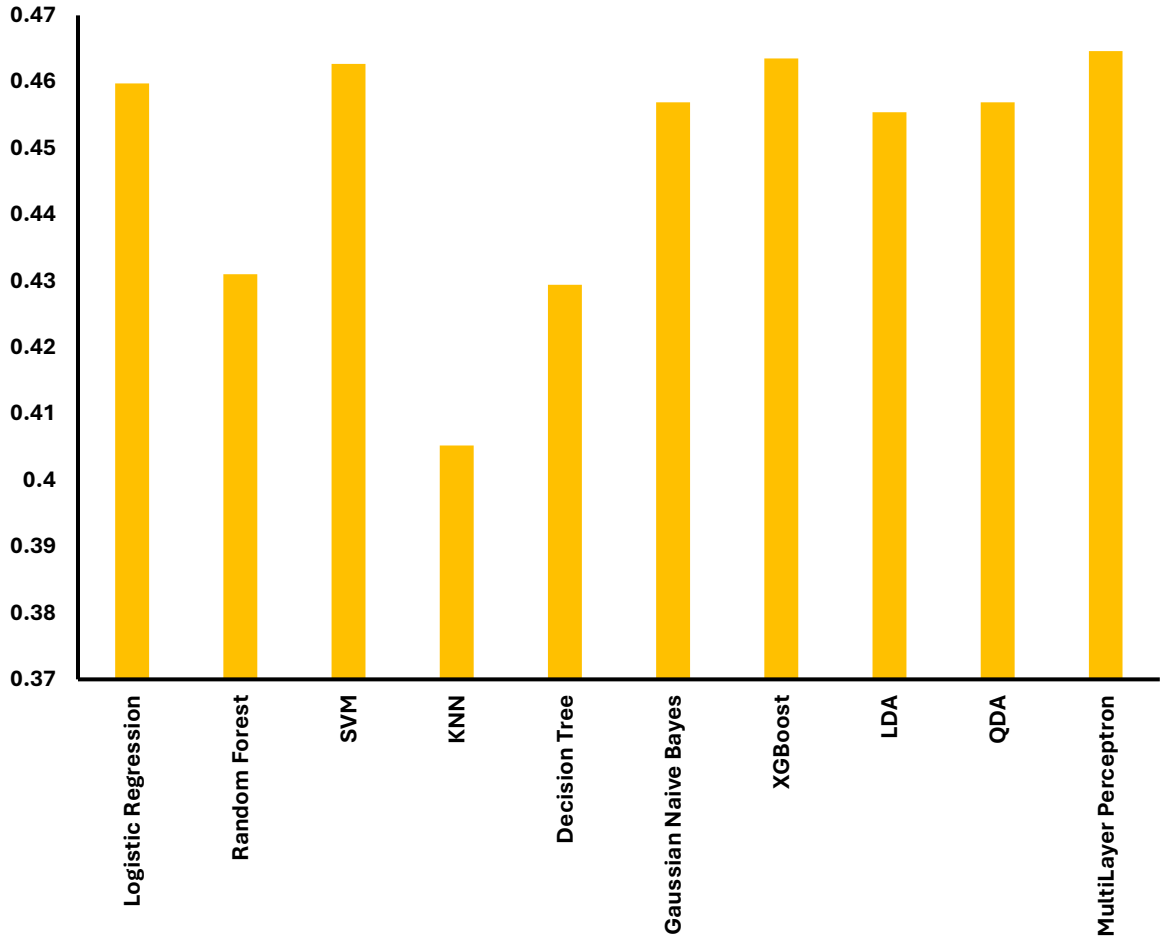

Figure:1. The X- axis represents the different machine learning models which were applied on the dataset and the Y- axis represents the accuracy scores of the different models.

Multi-Layer Perceptron, which is an ANN model was calculated as the best model in terms of accuracy, followed by XGBoost and SVM. The worst performing models were found to be KNN, Decision Tree and Random Forest as depicted in Figure:1.

Recall: It is known as sensitivity or true positive rate of a model. Recall measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (sum of true positives and false negatives). It focuses on the ability of the model to capture all positive instances avoiding the false negatives.

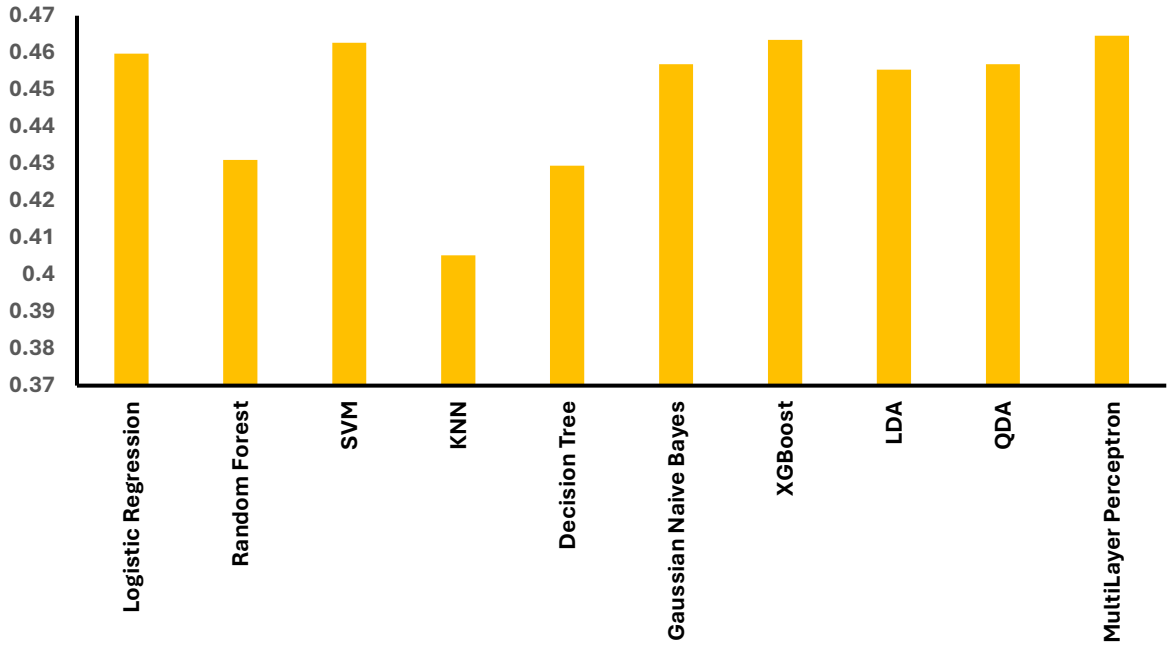

Figure:2. The X- axis represents the different machine learning models which were applied on the dataset and the Y- axis represents the recall values of the different models.

In this instance also, Multi-Layer Perceptron has the best recall value, followed by XGBoost and SVM. The models- KNN, Decision Tree and Random Forest were found to be the worst performing models in terms of their recall values.

Precision: It is calculated as the ratio of the true positives to the sum of true positives and false positives. This is one of the metrics of the classification models which indicates how many of the predicted positive instances are actually positive. Precision reflects the quality of the positive prediction made by the model.

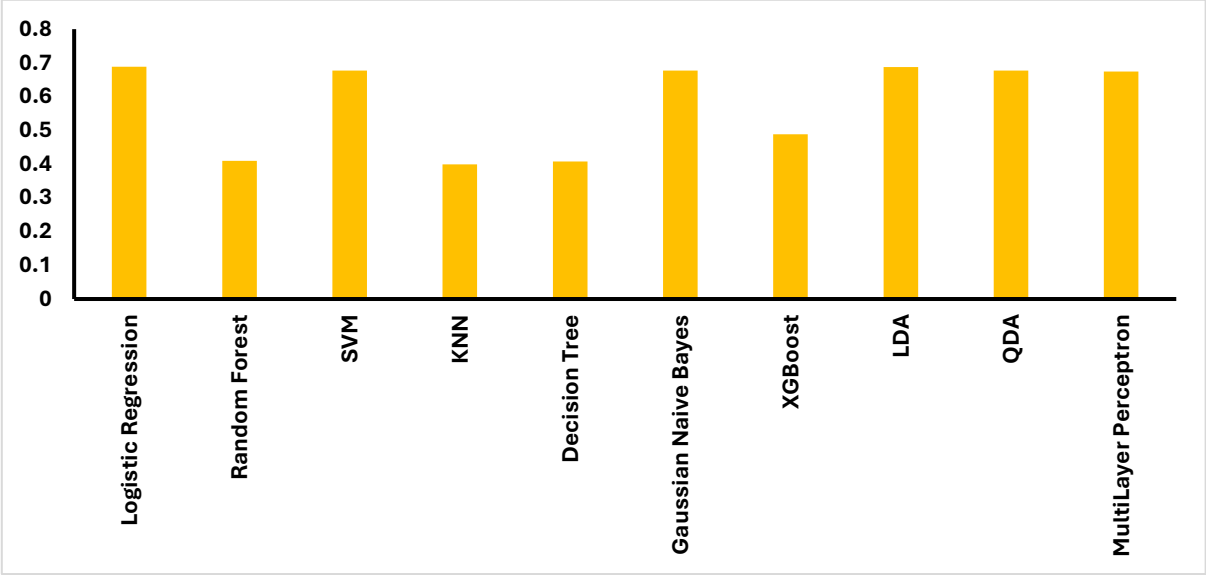

Figure:3. The X- axis represents the different machine learning models which were applied on the dataset and the Y- axis represents the precision scores of the different models.

Logistic Regression performs the best in terms of its precision values, followed by LDA and Gaussian Naïve Bayes and QDA, both having the same value. The worst performing models in terms of their precision were found to be KNN, Decision Tree and Random Forest.

F1-score: This is a very useful metric for measuring the performance of the classification models because it takes into account both false positive and false negative and not just the number of incorrect predictions. F1 score is calculated as the harmonic mean of precision and recall and gives a better understanding of the performance of the models.

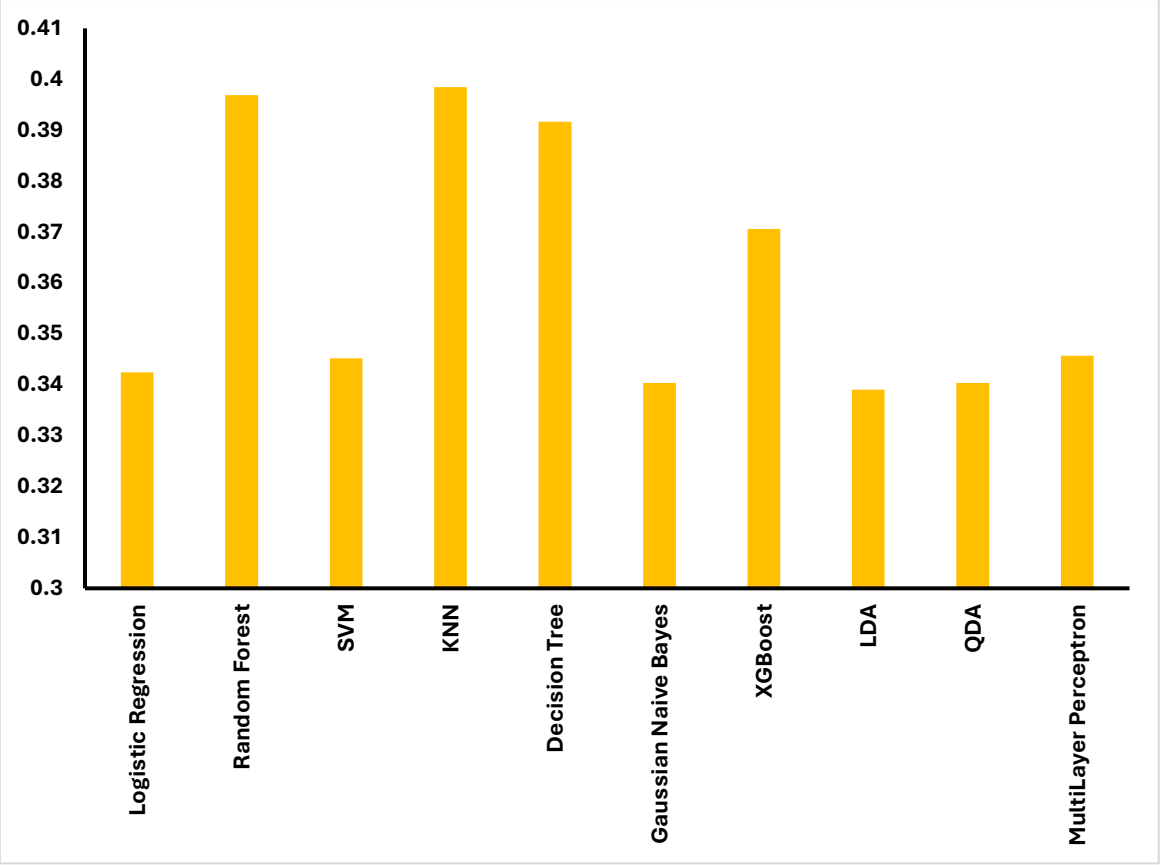

Figure:4. The X- axis represents the different machine learning models which were applied on the dataset and the Y- axis represents the F1 scores of the different models.

KNN performs the best in terms of F1 score, followed by Random Forest and Decision Tree. The worst performing models, by evaluation of their F1 scores were LDA. QDA and Gaussian Naïve Bayes were both found to have the same F1 scores.

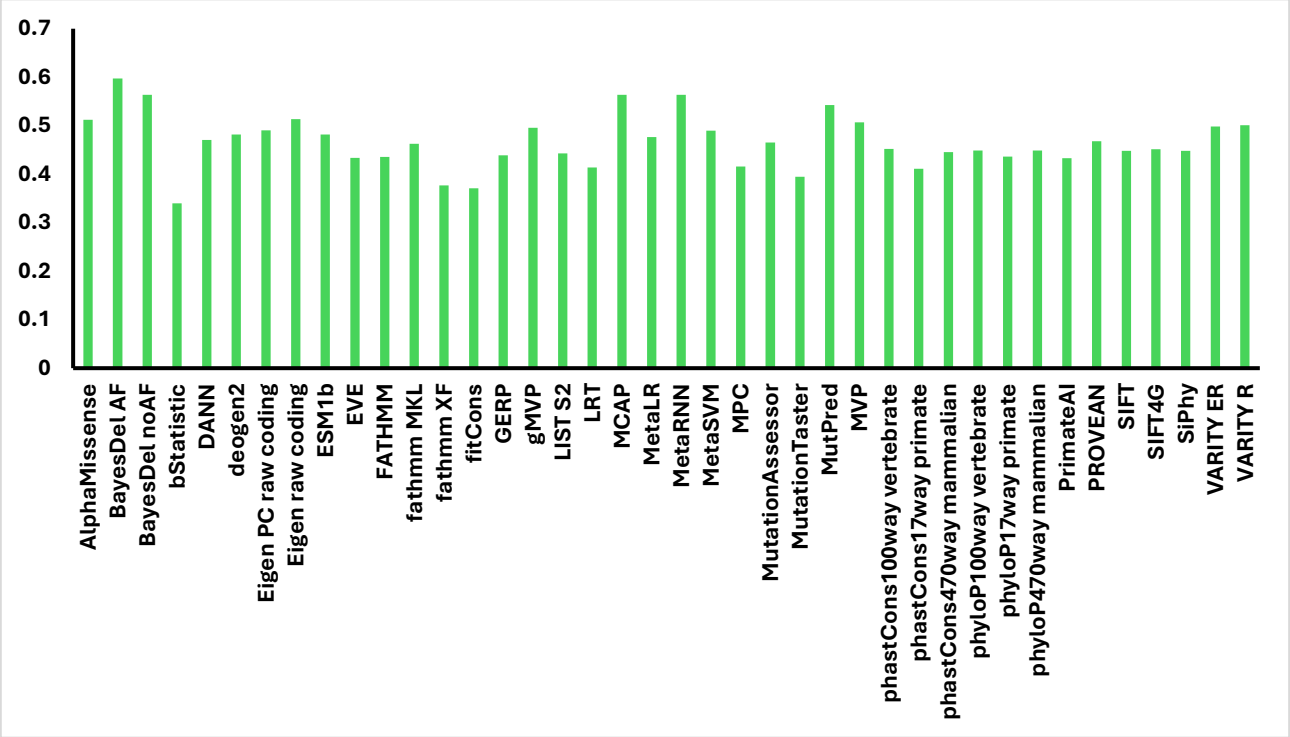

Figure:5. The X- axis represents the different prediction algorithms and conservation scores and the Y- axis represents their accuracy values. The median value of the accuracy scores predicted by each of the machine learning algorithms were considered.

Among the various functional prediction algorithms and conservation scores, the best performing algorithm was BayesDel AF, followed by BayesDel noAF, MetaRNN, MCAP and MutPred. The worst best performing instances were bStatistic, fitCons, fathmm XF, MutationTaster and phastCons17way primate.

Discussion:

This is a preliminary study to compare the different functional prediction algorithms and conservation scores in evaluating the clinical impact of the non-synonymous single nucleotide variants (nsSNVs) and splice site single nucleotide variants (ssSNVs) of the human genome. The study shows that no single instance can predict the pathogenicity or deleteriousness of a variant with full confidence. So, this observation tells us to use more than a single functional prediction algorithms or conservation scores while annotating the pathogenicity of a variant in a particular disease condition. The best performing features or instances can be prioritized over others while selecting the functional prediction algorithms, which are used to determine the driver mutations or the pathogenic variants from genomics or transcriptomics datasets. A combination of different algorithms would be the best way to find the pathogenic mutations from the omics datasets, instead of using a single algorithm.

Conclusion:

The same machine learning algorithms can generate different results across different study settings. A broader level classification of the algorithms to predict the clinical impact is also one of the limitations of this study. This study only attempted to compare the different functional prediction algorithms and

conservation scores to identify the best algorithms to be used for annotating a variant as pathogenic and having a deleterious effect on the patient.

Competing Interests:

The author declares no competing interests, whatsoever.

Availability of data and materials:

dbNSFPv4.7 database has been used in this study which can be downloaded from their website. (http://database.liulab.science/dbNSFP)

References:

1. Liu X, Jian X, and Boerwinkle E. 2011. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation. 32:894-899.
2. Liu X, Li C, Mou C, Dong Y, and Tu Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Medicine. 12:103.
3. Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8
4. T. M. Mitchell, "Machine learning WCB": McGraw-Hill Boston, MA:, 1997.
5. Ayer T, Chhatwal J, Alagoz O, Kahn CE Jr, Woods RW, Burnside ES. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13–22.
6. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018;132:1578–85.
7. Yang J, Yao D, Zhan X, Zhan X. Predicting disease risks using feature selection based on random forest and support vector machine. In: International Symposium on Bioinformatics Research and Applications; 2014. p. 1–11. Springer.
8. Tang Z-H, Liu J, Zeng F, Li Z, Yu X, Zhou L. Comparison of prediction model for cardiovascular autonomic dysfunction using artificial neural network and logistic regression analysis. PLoS One. 2013;8(8):e70571.
9. Marikani T, Shyamala K. Prediction of heart disease using supervised learning algorithms. Int J Comput Appl. 2017;165(5):41–4.
10. Lynch CM, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform. 2017;108:1–8.
11. Zupan B, DemšAr J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med. 2000;20(1):59–75.
12. Qorri, Erda et al. "A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants." International journal of molecular sciences vol. 23,14 7946. 19 Jul. 2022, doi:10.3390/ijms23147946
13. Thusberg J., Olatubosun A., Vihinen M. Performance of Mutation Pathogenicity Prediction Methods on Missense Variants. Hum. Mutat. 2011;32:358–368. doi: 10.1002/humu.21445.

14. Martelotto L.G., Ng C.K., De Filippo M.R., Zhang Y., Piscuoglio S., Lim R.S., Shen R., Norton L., Reis-Filho J.S., Weigelt B. Benchmarking Mutation Effect Prediction Algorithms Using Functionally Validated Cancer-Related Missense Mutations. Genome Biol. 2014;15:484. doi: 10.1186/s13059-014-0484-1
15. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011;12:2825–2830.
16. 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. Hum. Mol. Genet. 2015;24:2125–2137. doi: 10.1093/hmg/ddu733.
17. Ghosh R., Oak N., Plon S.E. Evaluation of in Silico Algorithms for Use with ACMG/AMP Clinical Variant Interpretation Guidelines. Genome Biol. 2017;18:225. doi: 10.1186/s13059-017-1353-5.