ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MACHINE LEARNING MODEL FOR MUTATION IMPACT PREDICTION BASED ON NETWORK PROPERTIES

BERK GÜRDAMAR

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Dr. Osman Uğur Sezerman

ISTANBUL-2022

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MACHINE LEARNING MODEL FOR MUTATION IMPACT PREDICTION BASED ON NETWORK PROPERTIES

BERK GÜRDAMAR

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Dr. Osman Uğur Sezerman

ISTANBUL-2022

Department:	Biostatistics and Bioinformatics
Program:	Biostatistics and Bioinformatics
Thesis Title:	Machine learning model for mutation impact prediction based on network properties
Student's name and Surname:	Berk Gürdamar
Date of Defence:	01/09/2022

This is to certify that I have examined this copy of master thesis. I have found that she/he prepared after fulfilling the specified requirements in the associated legislations before the final examining committee whose signatures are below.

Jury Member (Head of the Defense)	Assoc. Prof. Emel Timuçin Acıbadem University	Signature
Jury Member (Thesis Supervisor)	Prof. Dr. Uğur Sezerman Acıbadem University	Signature
Jury Member	Asst. Prof. Öznur Taştan Sabancı University	Signature

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethicel rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

Date

Berk Gürdamar

(Signature)

PREFACE AND ACKNOWLEDGEMENT

First, I would like to thank my supervisor, mentor, and idol Prof. Dr. Uğur Sezerman for his guidance, patience, and support over the years. It has been an honor to be a student of his. Also, I would like to thank to my jury members and substitutive jury members who are Assoc. Prof. Emel Timuçin, Asst. Prof. Öznur Taştan, Asst. Prof. Burcu Bakır Güngör and Prof. Dr. Eda Tahir Turanlı for their contribution.

Moreover, it has been an honor working with friends/colleges from Sezerman Lab. Especially, Ege Ülgen and Tuğçe Bozkurt for their help and support during the thesis study.

I would like to express my feelings to my family and close friends. They provided moral support while I have been working on my thesis. They have been very supportive in every case. Also, I would like to express most intense feeling to my girlfriend Ece Özlemiş. She is always with me in every condition, showed me her endless love, listened, and supported me in every complaint about life and work.

Lastly, I would like to thank European Joint Programme on Rare Diseases (EJP RD) for their monthly scholarship.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Å	Angstroms
Adaboost	Adaptive boosting
ASA	Accessible Surface Area
AUROC	Area under the receiver operating characteristic
BLAST	Basic Local Alignment Search Tool
CADD	Combined Annotation-Dependent Depletion
cryo-EM	Cryo-electron microscopy
Cα	Carbon alpha
DNA	Deoxyribonucleic acid
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
HGP	Human Genome Project
IDE	Integrated development environment
KEGG	Kyoto Encyclopedia of Genes and Genomes
LoF	Loss of Function
MAF	Minor allele frequency
NGS	Next-generation sequencing
NMR	Nuclear magnetic resonance
OGEE	Online GEne Essentiality
PDB	Protein Data Bank
pLI	Probability of loss of function intolerance
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SAVs	Single amino acid variants
SIFT	Sorting tolerant from intolerant
SNVs	Single nucleotide variations
TPM	Transcript per million
VUS	Variant of Uncertain Significance

WES	Whole exome sequencing
WGS	Whole genome sequencing
XGBoost	Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

ÖZET

Mutasyon Etkisi Tahmini İçin Ağ Özelliklerini İçeren Makine Öğrenme Modeli

Mutasyon etki tahmini biyoinformatik alanı için zorlu olmaya devam etmektedir. Mutasyonların patojenik veya nötr olarak sınıflandırılması, varyant önceliklendirme, varyantların daha iyi sınıflandırılması ve hastalıkların arkasındaki mekanizmanın daha iyi anlaşılması için önemli bir yaklaşımdır. Bu çalışmanın temel amacı, ağ özelliklerini kullanarak protein yapıları üzerinde mutasyon etkisi tahmini için bir makine öğrenme modeli oluşturmaktı. PDB yapıları 2 farklı atomlar arası etkileşim limiti kullanılarak ve 2 farklı ağ oluşum yaklaşımı kullanılarak ağ yapılarına dönüştürüldü ve 28 farklı makine öğrenme modeli geliştirildi. Ağ tabanlı, gen tabanlı ve amino asit tabanlı olmak üzere 3 alt kategoriye ayrılan 24 farklı özellik, daha iyi tahminler elde etmek için test edildi. Model oluşturma ve test için VariBench ve ClinVar veri setleri, doğrulama ve mevcut 32 farklı yöntemle karşılaştırma için Missense3D veri seti kullanıldı. Metodumuz, Missense3D veri setindeki tahminleriyle 0,9413 alıcı işletim karakteristiği altındaki alan değeri elde ederek diğer 32 metodun hepsinden daha iyi performans gösterdi. predatoR adında, en yüksek performans gösteren 7 angstrom mesafe limiti kullanılan ve bütün atomlar kullanılarak kurulan Adaboost modeli ve 5 angstrom mesafe limiti kullanılan ve sadece karbon alfa atomları kullanılarak kurulan Adaboost modelini içeren bir R paketi oluşturduk. predatoR aynı zamanda girdi veri setine 24 farklı özelliği hesaplamak ve anote etmek için 19 farklı fonksiyon içermektedir. predatoR, topluluğun kullanması için GitHub'da (https://github.com/berkgurdamar/predatoR) mevcuttur.

Anahtar Sözcükler: Mutasyon etkisi tahmini, Varyant sınıflandırması ve önceliklendirmesi, Ağ biçimlendirmesi, Makine öğrenmesi, predatoR R paketi

ABSTRACT

Machine Learning Model for Mutation Impact Prediction Based on Network Properties

Mutation impact prediction remains challenging for bioinformatics field. Classification of mutations as pathogenic or neutral is an important approach for variant prioritization, better classification of variants and better understanding of the mechanism behind diseases. Main purpose of this study was to build a machine learning model for mutation impact prediction on protein structures by using network properties. We developed 28 different machine learning models by converting PDB structures into networks with using 2 different interatomic interaction cutoffs and 2 different network formalization approach. 24 different features which divided into 3 sub-categories, network-based, gene-based and amino acid-based, were tested for getting better predictions. VariBench and ClinVar datasets were used for model building and testing, Missense3D dataset was used for validation and comparison with currently available 32 different methods. Our method outperformed all the other 32 methods with their predictions on Missense3D dataset with an area under the receiver operating characteristics curve value of 0,9413. We built an R package, predatoR, that contains the highest performed two models which were built with 7 angstrom distance cutoff and all atoms used Adaboost model and 5 angstrom distance cutoff and carbon alpha atoms used Adaboost model. predatoR contains 19 different functions for calculating and annotating 24 different features to the input dataset. predatoR is available on GitHub (https://github.com/berkgurdamar/predatoR) for community to use.

Keywords: Mutation impact prediction, Variant classification and prioritization, Network formalization, Machine learning, predatoR R package

1 INTRODUCTION

1.1 Personalized Medicine

In times past, medicine was based on signs and symptoms presented on a patient (1,2). Over the past decades, with the great effort on sequencing human genome started by Human Genome Project (HGP) (3) and developments in the next-generation sequencing (NGS) technologies, classical medicine evolved to personalized medicine (2,4,5). In personalized medicine, method goes beyond from the signs and symptoms to individual manner such as using information from lifestyle, clinical, genetic, health history and environment of an individual patient (2,6). Personalized medicine is widely used in the studies for better treatment opportunities for diseases such as cystic fibrosis, Alzheimer disease and different type of cancers (7–9).

Developments in the NGS technologies allowed us to access large amount of sequencing data. HGP was the pioneer of the sequencing studies. After the great effort of HGP, other projects such as 1000 Genomes Project (10), HapMap project (11) and Human Variome Project (12) started to produce genomic datasets. After the sequencing studies, databases started to developed for containing results from genome-wide sequencing studies.

With using datasets and databases, variations in the human genome can be identified. Mutations in the genome can cause change in the amino acid sequence of a protein. These alterations can change the protein folding, protein stability, interaction of active site and protein expression. As a result of these changes, structure or function of a protein can be affected (13). Non-synonymous single nucleotide variations (SNVs) are one type of mutations that cause amino acid change on the protein sequence. Non-synonymous SNVs are essential for identification of the variations' effect on human genome (14). This mutation type also called missense variants or single amino acid variants (SAVs) and they are the most studied group of variations (15).

Rare diseases defined as diseases that affects low number of individuals are caused by genetic variations. Definition of a rare disease is changing across the countries. In United States, a disease classified as rare if affects less than 200,000 individuals. In European Union, a disease seen in 1 in 2,000 individuals classified as rare diseases (16,17). With the developments in the NGS technologies, doctors are suggesting sequencing to patients which may have rare diseases (16). However, diagnostics still difficult even with the developments in the sequencing technologies. Many variants are classified as Variant of Uncertain Significance (VUS) which are mutations that there is limited or conflicting information about the effect of the mutation (18,19). Computational methods are necessary for classifying mutations especially classified as VUS and diagnostics of rare diseases become faster with the filtration of VUS variants with using computational methods.

1.2 History of Mutation Impact Prediction

Over the past decades, different approaches have been made for mutation impact prediction such as sequence conservation-based, structure-based, combined approaches which include sequence and structure together and meta-predictors which make prediction based on the predictions from other impact prediction methods (20,21).

First milestones were achieved by Miller and Kumar and they showed the power of sequence conversation analysis on the classification of variations as pathogenic or neutral (21,22). Importance of structural information was shown by Wang and Moult and the combined power on the mutation classification was shown by Chasman and Adams (21,23,24).

First approaches of sequence conservation methods were mainly based on probabilities of amino acid changes over the evolutionary time. Probabilities of amino acid substitutions were calculated by aligning homologous proteins' sequences. This substitution matrices became popular and used by algorithms such as Basic Local Alignment Search Tool (BLAST) (25) and FASTA (21,26). With the increase in the

sequence databases, position-specific probabilities started to be computed and performed better than the substitution matrices (21). Sorting tolerant from intolerant (SIFT) (27) is an example of widely used methods. SIFT uses conservation data, creates normalized probability matrix and make prediction by looking at the observed and the unobserved distribution of the amino acids (20,28).

Structure-based methods mainly focused on stability of a protein by identifying the difference in the free energy. If a mutation has a large effect on the stability of a protein, mutation most probably have an effect on the stability and function. Most of the prediction methods have energy-based models that calculates folded and unfolded states' energy of the protein. For classifying mutations, machine learning-based approaches have also been made for structure-based prediction (21).

Sequence and structure were combined by the prediction tools such as PolyPhen (29). PolyPhen uses Naïve Bayes Classifier for calculating impact of a mutation and uses features such as conservation data, functional domains of proteins and structural features (20). Machine learning methods dominates the impact prediction methods because of the flexibility of the algorithms such as including different type of features and have an improved performance over the non-machine learning-based methods (21). PrimateAI (30) is another example of impact prediction methods that uses deep neural network method with combining sequence and the secondary structure of a protein (28). Rhapsody (31) uses sequence-based, structure-based and dynamic-based features and predicts the impact of a mutation using a random forest model.

Meta-predictors started to develop with the increase in the impact prediction tools. Main aim of meta-prediction tools is using predictions of different methods for calculating better predictions. First approach was done by CONDEL (32) which uses predictions of 5 different prediction methods (21). Another example is The Combined Annotation-Dependent Depletion (CADD) (33) that uses a support vector machine model with 63 different features for calculating the prediction (28). REVEL (34) is an ensemble method that uses 13 different impact prediction scores of prediction tools

such as PROVEAN (35), FATHMM (36) and GERP++ (28,37). Many approaches also have been done on meta-prediction methods such as CoVEC (38), Meta-SNP (39) and MetaSVM (40).

1.3 Scientific Background

In this section, information about databases, machine learning algorithms and network properties used in this thesis study are given.

1.3.1 Databases

1.3.1.1 VariBench

VariBench (41) is a database that contains benchmarking datasets for variations. Datasets of VariBench chosen from literature, databases and contains experimentally verified variations. Different type of datasets were included such as variations mapped on protein, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and three-dimensional protein structures. Database is divided into 5 different categories, tolerance datasets, protein stability datasets, mismatch repair gene variants, transcription factor binding site dataset and splice site variation dataset (41).

1.3.1.2 ClinVar

ClinVar (42) is a publicly available archive that contains data about relationship between variant and human phenotype. Variations from clinical test, literature or research can be submitted to ClinVar (42). ClinVar uses dbSNP (43) and dbVar (44) databases for collecting information about variant location from human assemblies. Every submission has its own unique identifier in the database. ClinVar aggregates variants or phenotypes which has same combination and reports that if there is a conflicting information or not (42).

1.3.1.3 Missense3D

Missense3D (45) database contains datasets of human mutations. Missense3D is also a web server that predicts the change in the protein structure caused by an amino acid substitution with using Phyre2 homology model predictor (45). Experimental and predicted PDB structures are also available on Missense3D web server (http://missense3d.bc.ic.ac.uk/missense3d/index.html).

1.3.1.4 Protein Data Bank

The Protein Data Bank (PDB) (46) is an archive that contains structural information of biological molecules. Structures in the PDB are gathered from cryo-electron microscopy (cryo-EM), nuclear magnetic resonance (NMR), X-ray crystallography and theoretical modeling. PDB data contains sequence, coordinates of the atoms in the structure, cofactors, prosthetic groups, names of all components, description of structure and many information about the molecule (46).

1.3.2 Machine learning algorithms

Machine learning is a scientific discipline that based on computers. Main approach of machine learning algorithms is learning the data with identifying patterns from large datasets (47,48). Machine learning can be divided into 2 categories: Supervised learning and unsupervised learning. In supervised learning, labels are used for training the algorithm. On the other hand, in unsupervised learning, there is no labels used for training the algorithm and the algorithm tries to find patterns that separates the dataset (48,49). Tasks of machine learning algorithms can be divided into 2 groups: Classification and regression. Outputs of classification tasks are labels or categories that algorithm tries to predict the class of the data. In regression tasks, outputs are continuous variables such as predicting the height from weight (49).

1.3.2.1 Random Forest

Random Forest algorithm creates decision trees with using different part of the dataset. New sample goes through all the decision trees and final outcome of the algorithm created according to the results of tested decision trees (47,49,50). Deeper trees can cause overfitting but overfitting prevented by using all trees for the final decision (47).

1.3.2.2 XGBoost

Extreme Gradient Boosting (XGBoost) is an example of tree-based machine learning algorithms that generates trees and decides the final model based on the generated trees with using gradient boosting. L1 and L2 regularizations are used for reducing the over-fitting and increasing the generalization of the model (49,51).

1.3.2.3 Adaboost

Adaptive boosting (Adaboost) is another tree-based machine learning algorithm that uses weights on the samples for classification. In initial step, all weights are the same for samples but in the next steps of the algorithm, weights of the misclassified samples are increased, and the algorithm forced to classify them better (52–54). Adaboost is good for binary classification problems but noise of the data can affect the performance of the algorithm (49).

1.3.2.4 Naïve Bayes

Naïve Bayes algorithm is derived from Bayes' theorem. Algorithm accepts each feature independent from each other (55). The algorithm can be used for two class or multi-class classifications and can be trained with small sized dataset. However, assumption of independent variables can affect the performance of the algorithm (49).

1.3.2.5 GLM

Generalized Linear Model (GLM) formulized by John Nelder and Robert Wedderburn includes Linear regression, Logistic regression and Poisson regression (56). With using GLM, linear relationship can be created between the input and the output variable though their relationship is not linear. Moreover, GLM can be used with categorical output.

1.3.2.6 GBM

Gradient Boosting Machine (GBM) is a machine learning algorithm that uses boosting for increasing the model performance. GBM can be used for both regression and classification tasks (57). In GBM algorithm, models are created sequentially and each new model tries to decrease the error of the last model (58).

1.3.2.7 Elastic Net

Elastic Net algorithm combines both Lasso and Ridge regression methods. In Lasso regression, important predictors are selected and used in the model. In Ridge regression, less important predictors' coefficients set a number close to zero so that they have less effect on the model. Elastic Net uses L1 and L2 penalties from both regressions to improve the model performance (59,60).

1.3.3 Network properties

Eigen centrality also called eigenvector centrality is a measure for calculating the influence of a node in a network, in this theory, value of a node is directly related with its neighbors' value (61,62). Eigen centrality score is calculated by sum of the *degree centrality* scores of nodes that connected to a node. Degree centrality measured by calculating the number of links that a node has and it is a measure for identifying popularity of a node (62).

Betweenness centrality is a metric for calculating the importance of a node in a network. For calculating the Betweenness centrality, *Shortest Path centrality* also used. Shortest Path centrality defined as the shortest path distance between two nodes (63). Betweenness centrality measures all the shortest paths between nodes i, j when $i \neq j \neq k$ and calculates how many times the path go across the node k (62,63).

PageRank centrality algorithm introduced by Sergey Brin and Larry Page in 1998 and used as Google's page ranking algorithm (64). Basic idea behind the algorithm is ranking the pages according to their links. Important links receive more links from other websites (65).

Clique centrality measures the linked every two nodes in a network (63). In this thesis study, clique centrality computed for each node separately. Number of edges between neighbors of a node was calculated and used as a score of a node.

Network properties have been used for protein interaction identification but never used for mutation impact prediction (66). Network-based approach also have been made by ProSNEx (67) for analyzing protein structures as networks. Contact map networks are created from protein structures and further analysis can be done via ProSNEx. However, the tool can be used only exploratory purposes not for mutation impact prediction.

Main aim of this thesis study is developing a new machine learning-based method for mutation impact prediction. We built 28 different machine learning models with using 7 different algorithms, 2 interatomic interaction distance cutoffs and 2 network formalization approaches. We chose the best performed models with using predictions of machine learning models on Missense3D dataset and compare its performance with 32 different impact prediction method. With using final models, we built an R package called predatoR for scientific community to use our method for mutation impact prediction.

2 BACKGROUND

Signs and symptoms were used in the medicine before (1,2). But with the developments in the sequencing technologies and studies such as HGP and 1000 Genomes Project, classical medicine evolved into personalized medicine. Personalized medicine does not only use signs and symptoms, uses also personal information such as genetics and lifestyle. With personalized medicine, better diagnosis and better treatments can be done (2,4,5).

Sequencing studies produced large amount of sequencing data. Databases started to develop for containing those genomic datasets. From those datasets, mutations in the human genome can be identified and used for diagnosis and finding treatment opportunities. Effect of some mutations are known, but some of the variations' effect does not. Those variations are called VUS. With using sequencing data, mutations with unknown effects such as VUS can be classified using *in silico* tools. Sequence conservation, structure and combined approaches have been done over the past decades. Also, meta-prediction is another approach for mutation impact prediction. In meta-prediction, predictor uses other predictors' prediction for better classification (20,21).

Machine learning is a scientific discipline that its main approach is learning the dataset by identifying the patterns (47,48). Supervised and unsupervised learning are the 2 groups of machine learning algorithms. Labels for the dataset are used in supervised learning. In supervised learning, no labels are used (48,49). Tasks of machine learning algorithms are divided into 2 categories. Classification is the one category that the main approach of classification task is dividing data into groups with using labels. In regression tasks, continuous variables are the outputs of the algorithms (49). Random Forest, Linear Regression, and Logistic Regression are the examples of machine learning algorithms.

Network properties have been used before for predicting protein interactions (66). Also, network-based approach have been made for converting protein structures into

networks by ProSNEx (67). ProSNEx uses contact map networks for protein structure analysis, but the tool does not offer any impact prediction, it can be used for only structure exploration.

In this thesis study, we developed a new machine learning-based method for mutation impact predictions. ClinVar and VariBench datasets were used for building machine learning models using 7 different machine learning algorithms and 24 unique features. Missense3D dataset was used for model validation and comparison with currently available 32 different method. Our method outperformed all 32 other prediction methods. We developed an R package called predatoR that contains the best performed 2 models and 19 different functions for community to use.

3 MATERIALS AND METHODS

All analyzes were done by using R v4.1.0 (68) with using an integrated development environment (IDE) RStudio 2021.09.0+351 "Ghost Orchid" Release (69). Overall workflow of the study represented in the Figure 1.

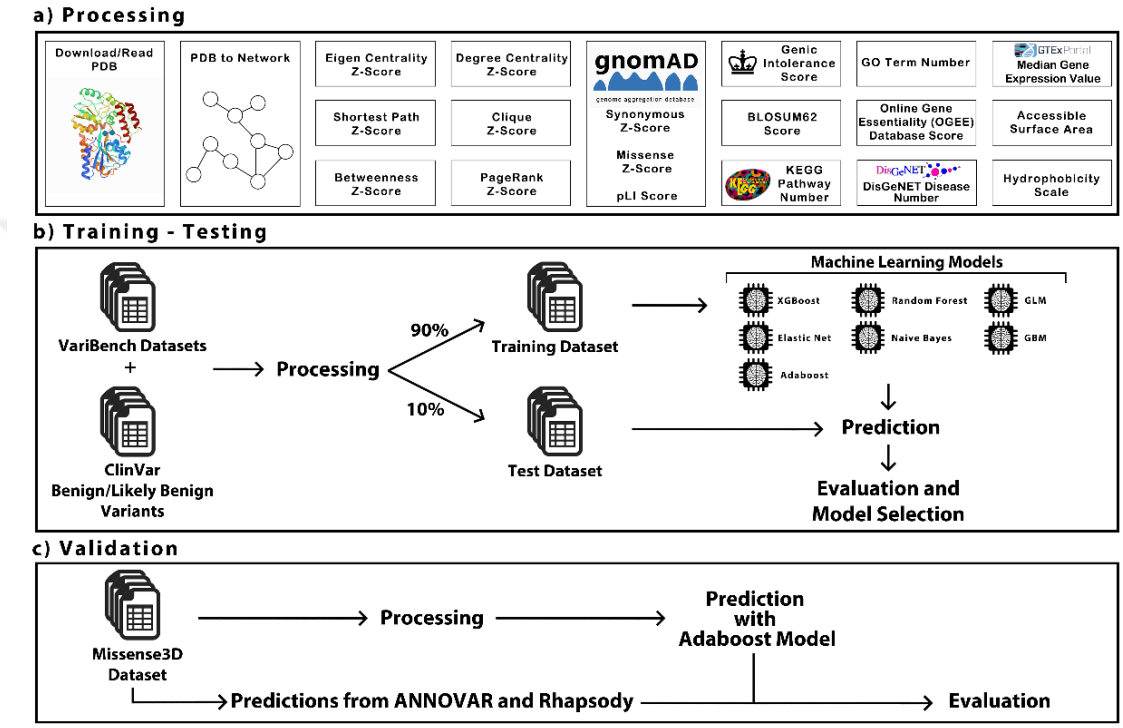

Figure 1. Overall workflow of the study. (A) All processing steps that were done on each mutation. (B) Training, testing, model evaluation and selection steps that were done on VariBench and ClinVar datasets. (C) Validation and comparison between our method and currently available mutation impact prediction methods by using Missense3D dataset.

3.1 Dataset Collection

VariBench datasets were selected to use as training and testing datasets for building machine learning models. Neutral (DS2, DS4, DS6, DS8, DS10, DS12, DS14, DS16, MMR and dbSNP), pathogenic (DS3 and DS5) and training dataset of

Rhapsody were collected from VariBench web server and datasets were combined. Four letter PDB ID containing variations were filtered, duplicated mutations and mislabeled mutations which means different labels (pathogenic and neutral) for the same mutation removed. Combined VariBench dataset contained 6,983 neutral and 23,563 pathogenic mutations. For increasing the number of neutral variants and preventing a class imbalance, dataset from ClinVar were downloaded. ClinVar dataset contained more than 1,000,000 unique variations. Benign and likely benign variants were filtered from the dataset. Variations were mapped on only genomic positions, for converting to PDB coordinates, VarMap (70) was used. VarMap is a web-based tool that provides coordinate conversion between genomics coordinates to protein structure and sequence. From the VarMap results, only the variants mapped on PDB structures were filtered and combined with the VariBench dataset. Final dataset contained 48,005 mutations (24,496 neutral and 23,509 pathogenic variants). Gene name information of some variants were not included in the dataset. For assigning the missing gene name information, Ensemble BioMart (71) web server was used for assigning PDB-chain IDs to the associated gene names. If there are multiple genes associated with the same PDB-chain ID, higher gnomAD metrics containing gene was selected.

Two experimental datasets obtained from Missense3D webserver (http://missense3d.bc.ic.ac.uk/missense3d/index.html). Combined dataset contained 10,229 unique mutations (4,652 neutral, 5,577 pathogenic variants) mapped on UniProt and PDB IDs. Positions of the variations were mapped on only UniProt position, for converting the UniProt-based positions to PDB-based positions, PDBSWS (72) dataset were used. Common variants between VariBench-ClinVar dataset and Missense3D dataset were removed for preventing a bias, and the final size of the dataset decreased to 6,554 mutations (3,569 neutral, 2,985 pathogenic variants).

3.2 Network Formalization from PDB Structure

PDB structures in the datasets were downloaded with using Bio3D (73) R package. For converting a PDB structure into a network, we used 2 different approaches, first one is using all atoms in the structure, and the other one is using only

the carbon alpha (Cα) atoms. For these approaches, only structure related atoms were filtered, and water molecules were removed. According to the approach, all atoms or only Cα atoms were filtered and distances between each atom were calculated. In our approach, atoms are accepted as the nodes and relationships between atoms according to distance cutoffs are the edges of the network. We used two different interatomic interaction distance cutoffs for building networks, 5 and 7 angstroms (Å). After calculating all the distances between each atom, we set an edge between atoms which were interacted within distance cutoffs. From the calculated list of interactions, networks were created for each chain separately with using igraph (74) R package.

3.3 Features for Machine Learning Model Building

Twenty-four different unique features were used for model building. Features divided into 3 sub-categories: Network-based, gene-based, amino acid-based. Overall workflow is shown in Figure 2.

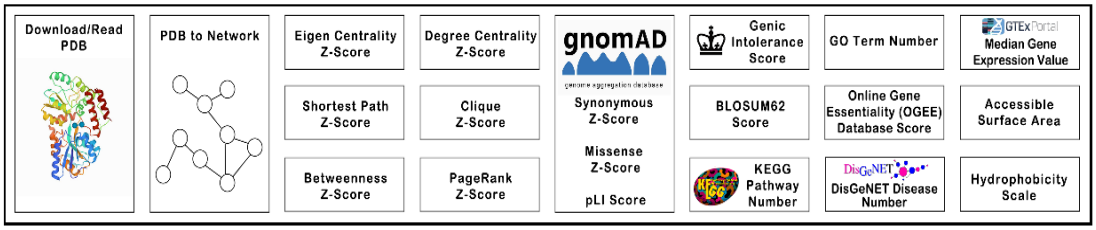

Figure 2. Overall processing steps of the study. Each PDB structure was downloaded and turned into a network. Network-based, gene-based and amino acid-based 24 features were calculated and assigned separately.

3.3.1 Network-based features

Six different network-based features were used for building machine learning models, Eigen Centrality, Betweenness Centrality, Degree Centrality, PageRank Centrality, Shortest Path Centrality and Clique Centrality. Except Clique Centrality, all the other features were calculated with using igraph R package. Clique Centrality was calculated by in house scripting. Each feature except Shortest Path Centrality gave

a single score for a node, Shortest Path Centrality gave all the shortest path distances of a node to each node. For calculating a single score from path lengths, all path lengths were summed. For Eigen Centrality, number of interactions between neighbors of a node was calculated. All scores of network-based features were turned into a Z-score for normalizing the scores between different PDB structures. For each mutation, Cα atoms' score were used as scores of a mutation for each network-based feature.

3.3.2 Gene-based features

Nine different gene-based features were used for building machine learning model, Missense Z-scores, Synonymous Z-scores and probability of loss of function intolerance (pLI) scores from gnomAD (75), number of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (76) and Gene Ontology (GO) (77) terms associated with input genes, associated disease number from DisGeNET (78), Genic Intolerance Score (79), gene essentiality score from Online GEne Essentiality (OGEE) v3 (80) database and median gene expression values of genes from Genotype-Tissue Expression (GTEx) portal (81).

3.3.2.1 Metrics from gnomAD

gnomAD is a database which aggregated from sequencing of more than 125,000 human whole exome sequencing (WES) and 15,000 human whole genome sequencing (WGS) data (75). Constrain metrics of gnomAD database are based on observed vs. expected number of rare SNVs of a gene. Missense Z-scores, Synonymous Z-scores and pLI scores were used as features in machine learning models. Higher Z-scores mean lesser number of variants were observed than the expected. pLI score is another gnomAD metric that calculates the likelihood of falling into the class of Loss of Function (LoF)-haploinsufficiency of a gene. Higher pLI scores indicates less tolerance to variation (75,82).

3.3.2.2 Associated KEGG pathway number

KEGG database consist of 15 both manually and computationally curated databases (76). All human KEGG pathways were downloaded with their associated gene information by using EnrichmentBrowser (83) R package. Gene symbols in the dataset were converted to HGNC gene symbols for calculating the associated KEGG pathway numbers for each gene in the dataset.

3.3.2.3 Associated GO term number

The aim of GO is creating a controlled language and apply it to all eukaryotes. GO terms divided into 3 categories, biological process, molecular function, and cellular component. Biological events a gene contributes referred with biological process category. Molecular function represents the activity of a gene product at molecular level. Place of a cell that a gene product shows activity represented by cellular component category (77). All human GO Terms with experiment evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, HAD, HMP and HEP) were collected from Ensemble BioMart. Biological process domain related GO terms filtered, and datasets were annotated according to the number of GO terms a gene was associated.

3.3.2.4 Associated disease number from DisGeNET

DisGeNET is a human database for gene-variant associations with diseases. curated datasets from experts, GWAS catalogue, literature and animal models were integrated for building DisGeNET datasets (78). Curated gene-disease associations dataset was downloaded from DisGeNET web server (https://www.disgenet.org/home/) and datasets annotated with number of diseases that a gene was associated.

3.3.2.5 Genic Intolerance score

Genic Intolerance uses more than 6,500 WES data for developing a score that based on observed and expected number of variations found on a gene (79). Genic Intolerance dataset was retrieved from their web server (https://genic-intolerance.org/) and scores based on minor allele frequency (MAF) filter at 0.1% were selected. Datasets were annotated according to gene names.

3.3.2.6 Gene essentiality scores from OGEE v3 database

OGEE v3 is an online gene essentiality database that contains datasets for 91 different species. Their method based on assigning genes as essential or not essential by analyzing gene functions and both environmental and experimental disturbances (80). OGEE v3 scores of input genes were used as gene essentiality scores.

3.3.2.7 Median gene expression value from GTEx Portal

GTEx Portal is a database of resources that contains tissue expression datasets for multiple tissues (81). Tissue-specific median transcript per million (TPM) gene expression dataset was downloaded from GTEx Portal web server (https://gtexportal.org/home/). Dataset contains median TPM gene expression values of more than 56,000 different genes for 54 different tissue types. Median gene expression values across 54 different tissue types were calculated for each gene. Datasets were annotated with the median gene expression values according to gene names.

3.3.3 Amino acid-based features

Hydrophobicity Scale (84) and Accessible Surface Area (ASA) (85) values were collected from aaSEA (86) R package. 3 different features were calculated and used from each property. Hydrophobicity Scale and ASA values of reference amino

acids, mutant amino acids and difference between reference and mutant amino acids were calculated and used as features.

BLOSUM62 is a matrix that contains scores for substitutions between each amino acid. BLOSUM62 matrix was introduced by Steven Henikoff and Jorja Henikoff and scores were calculated from alignment of more than 500 groups of proteins which were related (87). Substitution scores between reference and the mutant amino acids were used as another amino acid-based feature.

3.4 Dataset Preparation and Building of Machine Learning Models

Twenty-four different features were calculated for combined VariBench and ClinVar dataset. After calculations, the size of the dataset was reduced to 34,968 mutations (17,469 neutral, 17,499 pathogenic variants) because of missing information such as Genic Intolerance score of a gene. Dataset was randomly split into 90% training and 10% testing datasets with equally dividing the number of neutral and pathogenic variants.

Seven different machine learning algorithms were used, XGBoost, Naïve Bayes, Random Forest, Adaboost, Elastic Net, GLM and GBM. Twenty-eight different machine learning models were built, 7 models for 5Å distance cutoff and Cα only atoms, 7 models for 5Å distance cutoff and all atoms, 7 models for 7Å distance cutoff and Cα only atoms and 7 models for 7Å distance cutoff and all atoms. Parameter optimization was done through three repeated 10-fold cross validation for training all 28 machine learning models. All models were built with using same training and testing datasets. caret (88) R package was used for models training and testing. Model parameters can be found in the Table 1.

Table 1. Parameters of 7 different machine learning algorithms

Algorithm	Parameters
Adaboost	mfinal = 200, maxdepth = 18, coeflearn = "Breiman", method = "repeatedcv", number = 10, repeats = 3
GBM	interaction.depth = 12, n.trees = 200, shrinkage = 0.1, n.minobsinnode = 1, method = "repeatedcv", number = 10, repeats = 3
GLM	family= binomial, method = "repeatedcv", number = 10, repeats = 3
ElasticNet	tuneLength = 10, method = "repeatedcv", number = 10, repeats = 3
Random Forest	method = "repeatedcv", number = 10, repeats = 3
Naïve Bayes	method = "repeatedcv", number = 10, repeats = 3
XGBoost	nrounds = 500, max_depth = 13, eta = 0.05, gamma = 0.01, colsample_bytree = 0.75, min_child_weight = 0, subsample = 0.5, tuneLength = 1000, method = "repeatedcv", number = 10, repeats = 3

3.5 Validation, Model Selection and Comparison with Other Methods

Twenty-four features which were used in model building were also calculated for mutations in the Missense3D dataset. Same processing steps were done for 5Å-all atoms, 5Å-Cα atoms, 7Å-all atoms and 7Å-Cα atoms approaches. Predictions were calculated using 28 different machine learning models. Performance of 28 machine learning models were visualized by receiver operating characteristic (ROC) curve and area under the receiver operating characteristic (AUROC) curve values. Best model was chosen according to their AUROC curve values.

For method comparison, ANNOVAR (89) and Rhapsody were used. ANNOVAR is an annotation tool for genetic variants from sequencing data and can make gene-based, region-based and filter-based annotations. ANNOVAR gets input as genomic positions but, Missense3D dataset only contained the PDB coordinates of the variants. For getting genomic positions from PDB coordinates, TransVar (90) tool was used. TransVar is a multi-way annotator that can convert coordinates between genomic, transcript dependent cDNA and protein positions. Coordinate converted Missense3D dataset was annotated by using ANNOVAR with using “dbnsfp4.2a” database. 31 different prediction tools’ scores and predictions annotated to the dataset by

ANNOVAR. Also, predictions of Rhapsody on Missense3D dataset were also collected from Rhapsody web server (http://rhapsody.csb.pitt.edu/). Performance of prediction tools were visualized by ROC curves and AUROC curve values.

4 RESULTS

4.1 Results of Machine Learning Models

Twenty-eight different machine learning models were built with using 2 different interatomic interaction distance cutoffs and 2 network formalization approaches. According to the AUROC curve values on the test dataset (Figure 3. and Figure 4.), Adaboost models showed the best performance on both distance cutoffs.

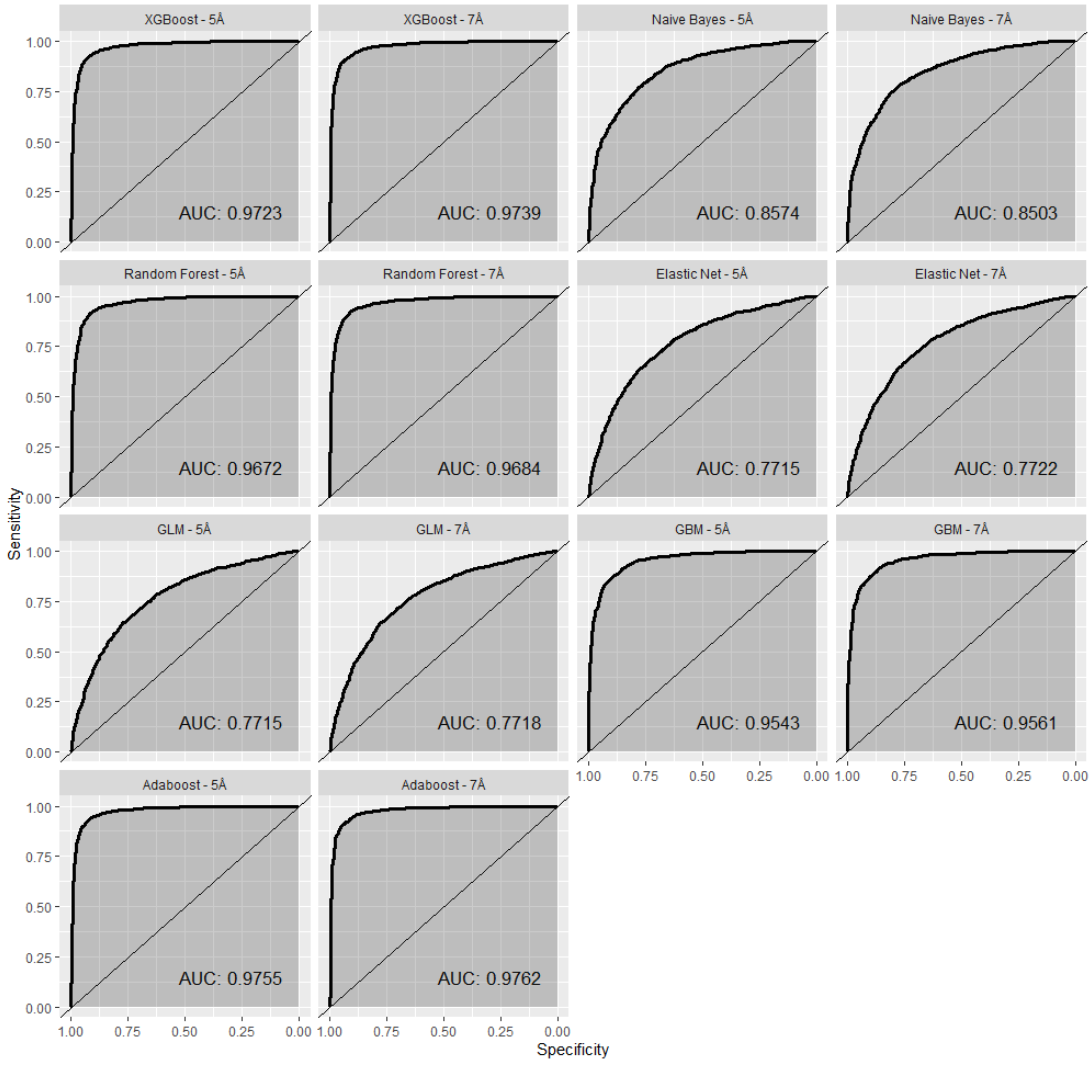

Figure 3. ROC curves of 14 machine learning models' predictions on the test dataset for both distance cutoffs. Models were built with using all atoms in the structure. Each plot was labelled with its AUROC curve values.

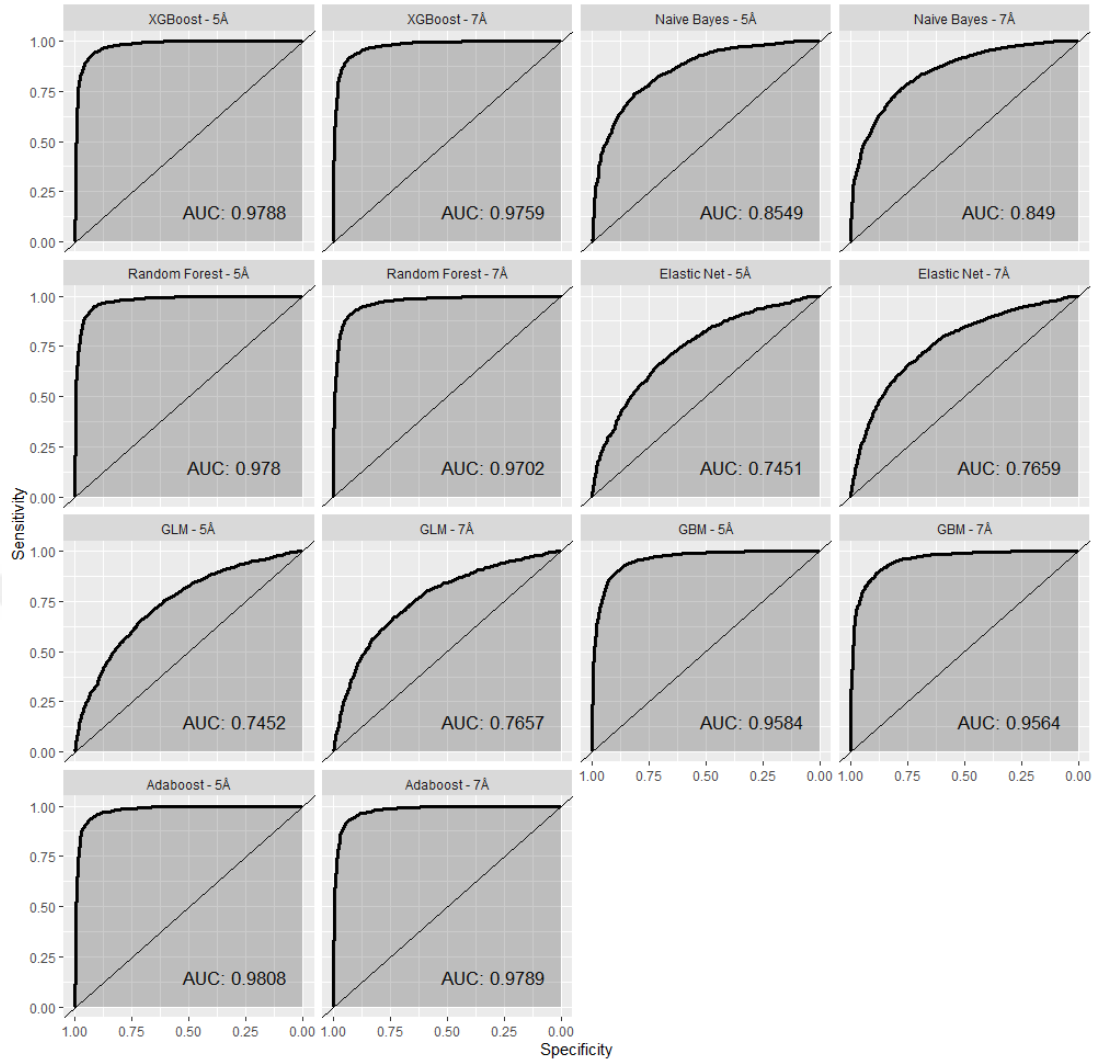

Figure 4. ROC curves of 14 machine learning models' predictions on the test dataset for both distance cutoffs. Models were built with using only Cα atoms in the structure. Each plot was labelled with its AUROC curve values.

Statistics such as accuracy, sensitivity, specificity values of 28 different machine learning models were calculated and shown in Table 2. As can be seen from the Table 2, Adaboost models showed the best performance for both 5Å and 7Å distance cutoffs and 2 network formalisation approaches.

Table 2. Accuracy, sensitivity, specificity, precision, recall, f1, prevalence and balanced accuracy values of machine learning models. Statistics are based on test dataset.

		Method	accuracy	sensitivity	specificity	precision	recall	f1	prevalence	Balanced accuracy
7Å	All atoms	Adaboost	0,923	0,923	0,922	0,922	0,923	0,923	0,500	0,923
		GBM	0,886	0,893	0,880	0,882	0,893	0,887	0,500	0,886
		GLM	0,708	0,703	0,713	0,710	0,703	0,706	0,500	0,708
		Elastic Net	0,710	0,703	0,717	0,713	0,703	0,708	0,500	0,710
		Random Forest	0,915	0,919	0,911	0,912	0,919	0,916	0,500	0,915
		Naive Bayes	0,751	0,626	0,876	0,835	0,626	0,716	0,500	0,751
		XGBoost	0,919	0,922	0,916	0,916	0,922	0,919	0,500	0,919
	Cα atoms	Adaboost	0,931	0,930	0,932	0,932	0,930	0,931	0,500	0,931
		GBM	0,888	0,891	0,884	0,885	0,891	0,888	0,500	0,888
		GLM	0,701	0,689	0,713	0,706	0,689	0,697	0,500	0,701
		Elastic Net	0,699	0,686	0,712	0,705	0,686	0,695	0,500	0,699
		Random Forest	0,917	0,922	0,913	0,914	0,922	0,918	0,500	0,917
		Naive Bayes	0,747	0,610	0,885	0,841	0,610	0,707	0,500	0,747
		XGBoost	0,924	0,923	0,925	0,925	0,923	0,924	0,500	0,924
5Å	All atoms	Adaboost	0,928	0,932	0,924	0,925	0,932	0,928	0,501	0,928
		GBM	0,881	0,887	0,874	0,876	0,887	0,881	0,501	0,881
		GLM	0,705	0,697	0,714	0,709	0,697	0,703	0,501	0,705
		Elastic Net	0,704	0,695	0,714	0,709	0,695	0,702	0,501	0,704
		Random Forest	0,914	0,918	0,909	0,910	0,918	0,914	0,501	0,914
		Naive Bayes	0,750	0,614	0,886	0,843	0,614	0,710	0,501	0,750
		XGBoost	0,923	0,928	0,919	0,920	0,928	0,924	0,501	0,923
	Cα atoms	Adaboost	0,934	0,937	0,931	0,933	0,937	0,935	0,506	0,934
		GBM	0,893	0,910	0,875	0,882	0,910	0,896	0,506	0,893
		GLM	0,681	0,667	0,695	0,692	0,667	0,679	0,506	0,681
		Elastic Net	0,683	0,668	0,698	0,694	0,668	0,681	0,506	0,683
		Random Forest	0,930	0,931	0,929	0,930	0,931	0,931	0,506	0,930
		Naive Bayes	0,750	0,610	0,892	0,853	0,610	0,712	0,506	0,751
		XGBoost	0,930	0,933	0,926	0,928	0,933	0,931	0,506	0,930

Three repeated 10-fold cross validation was used for building of each model. Mean accuracy values of cross validation results were also calculated and represented in Figure 5. and Figure 6. According to their mean accuracy values, best performed models were built with using Adaboost algorithm.

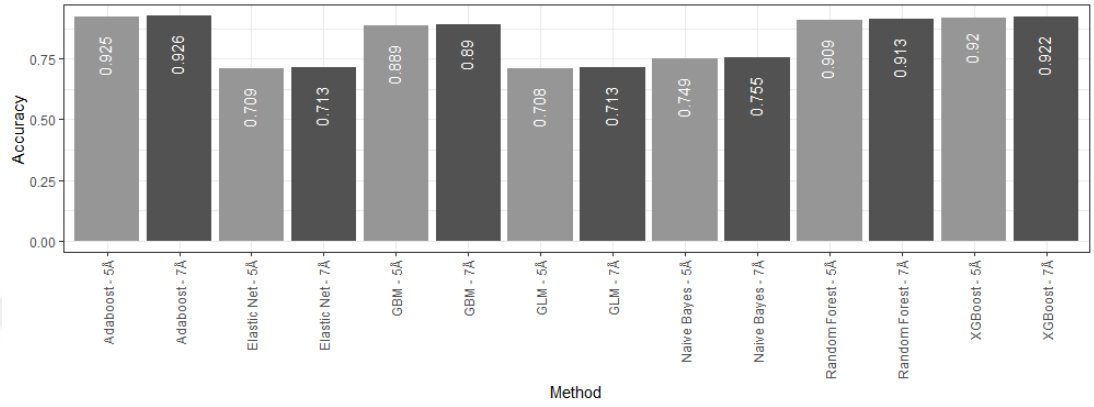

Figure 5. Mean accuracy values of cross validations of machine learning models built with all atoms in the structure.

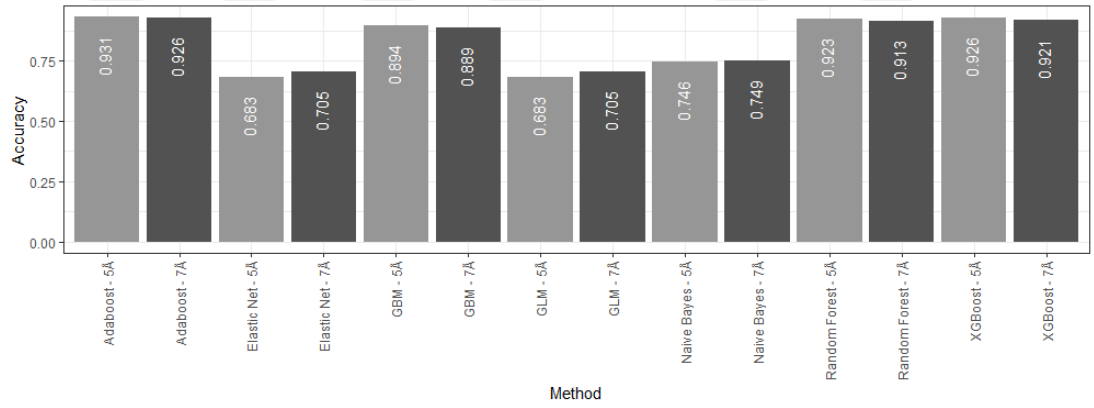

Figure 6. Mean accuracy values of cross validations of machine learning models built with Cα atoms in the structure.

According to the performances of models on test dataset, for all atom approach, Adaboost-7Å model showed highest AUROC curve value. For Cα atoms approach, Adaboost-5Å model showed highest AUROC curve value. These two models were

selected as final models and used for comparison with available 32 different impact prediction methods.

Performance of all machine learning models also assessed on Missense3D validation dataset by calculating AUROC curve values, demonstrated in Figure 7. and Figure 8.

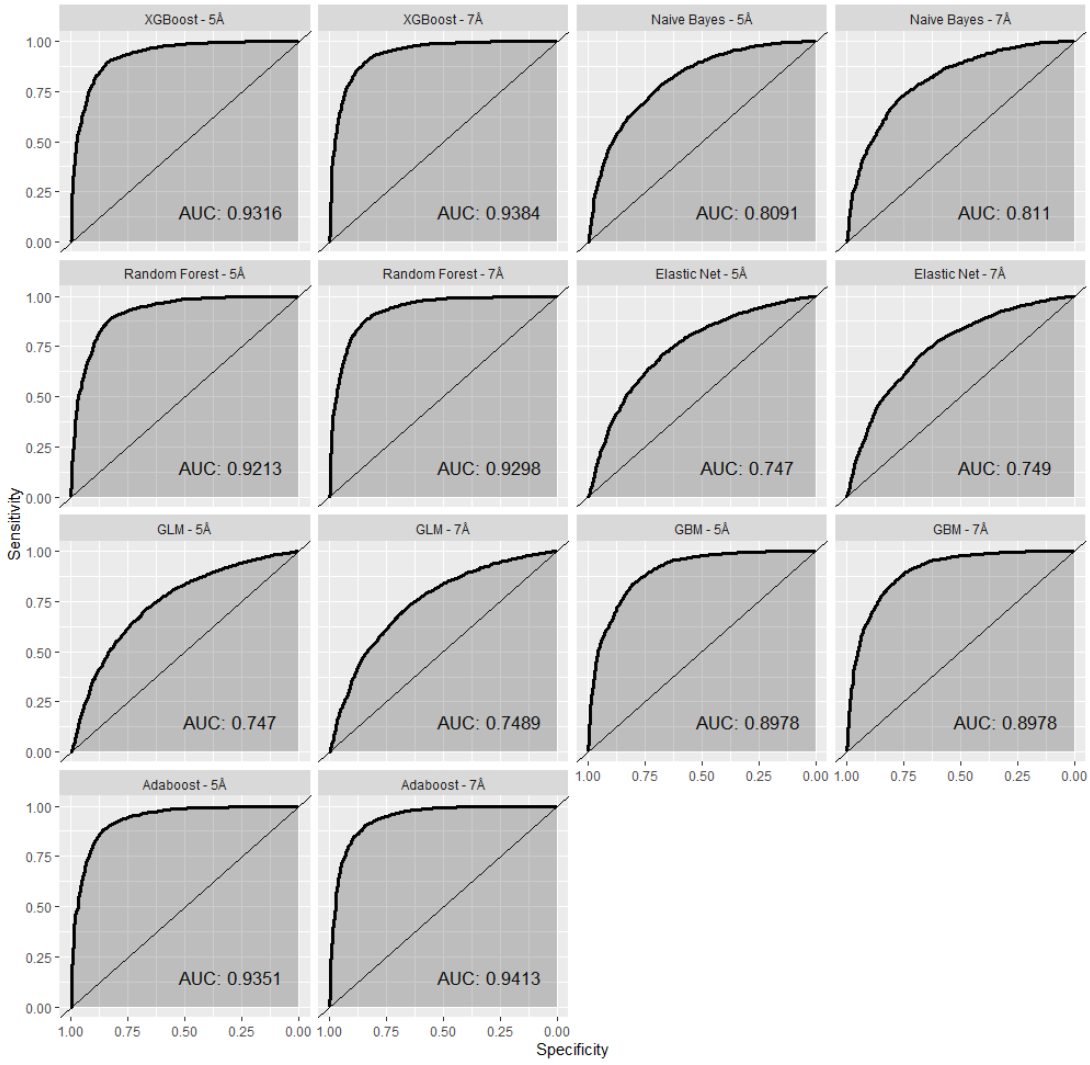

Figure 7. ROC curves of 14 machine learning models' predictions on validation dataset for both distance cutoffs. Networks were created by using all atoms in the structure. Each plot was labelled with its AUROC curve values.

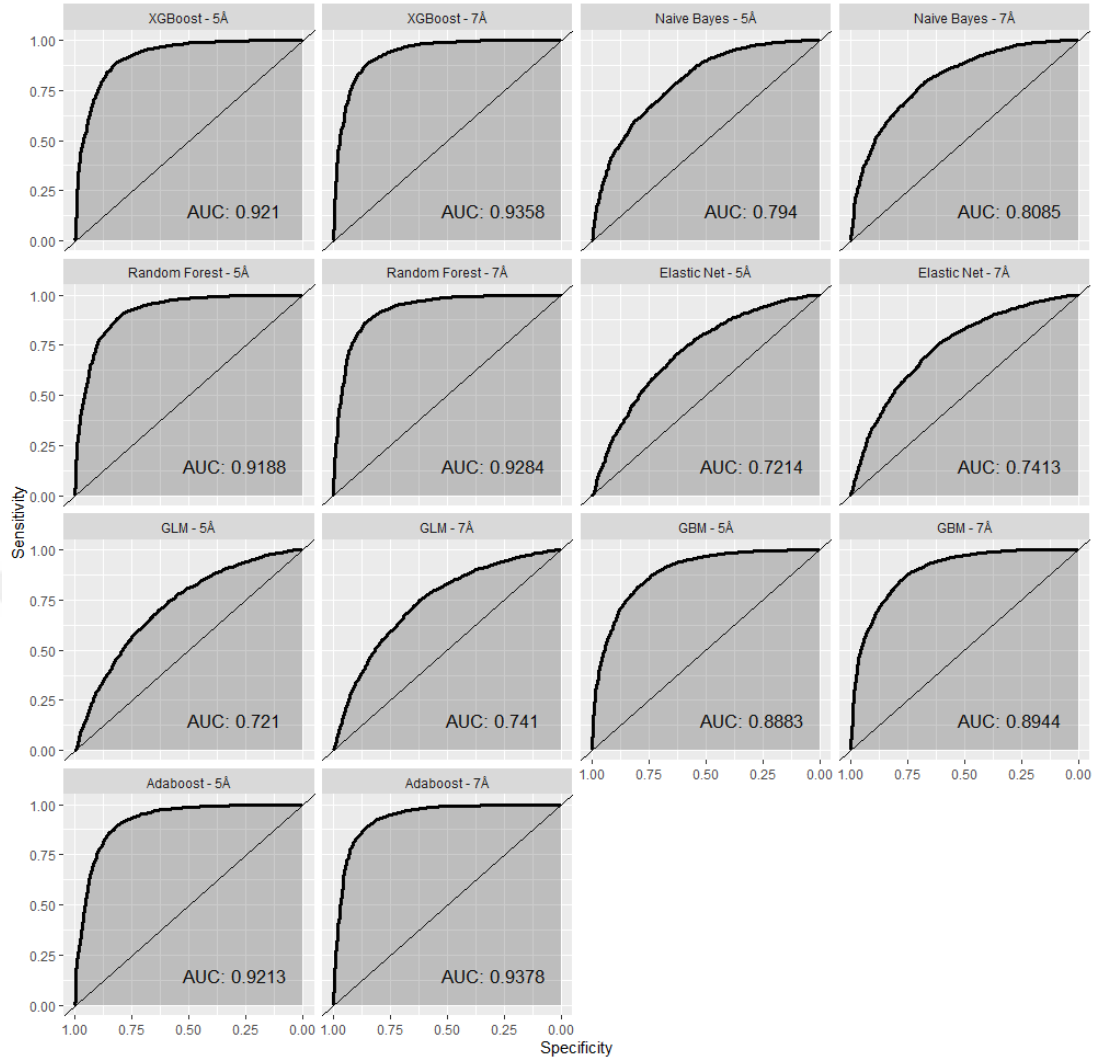

Figure 8. ROC curves of 14 machine learning models' predictions on validation dataset for both distance cutoffs. Networks were created by using Cα atoms in the structure. Each plot was labelled with its AUROC curve values.

4.2 Comparison with Other Methods

Prediction results of 31 different methods (LRT, SIFT, SIFT4G, Polyphen2_HDIV, Polyphen2_HVAR, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST4, MetaSVM, MetaLR, MetaRNN, M.CAP, REVEL, MutPred, MVP, MPC, PrimateAI, DEOGEN2, BayesDel_addAF, BayesDel_noAF, ClinPred, LIST.S2, DANN, fathmm.MKL_coding, fathmm.XF_coding, GenoCanyon, integrated_fitCons, GERP and CADD) were annotated to the Missense3D dataset by

ANNOVAR. ANNOVAR output was contained two different scores for predictions, “score” and “rankscore”. For comparison, “score” containing column names were filtered except LRT tool score. “LRT_converted_rankscore” column used as LRT score. “classi” column from Rhapsody prediction output contained the probabilities, therefore probabilities in that column used for comparison. Performance of all methods visualized by ROC curve and AUROC curve values (Figure 9.). As shown in the Figure 9., predatoR outperformed 32 other impact predictions tools with an AUROC curve value of 0.941 with 7Å-all atoms approach. Feature importance of the final models were shown in the Figure 10.

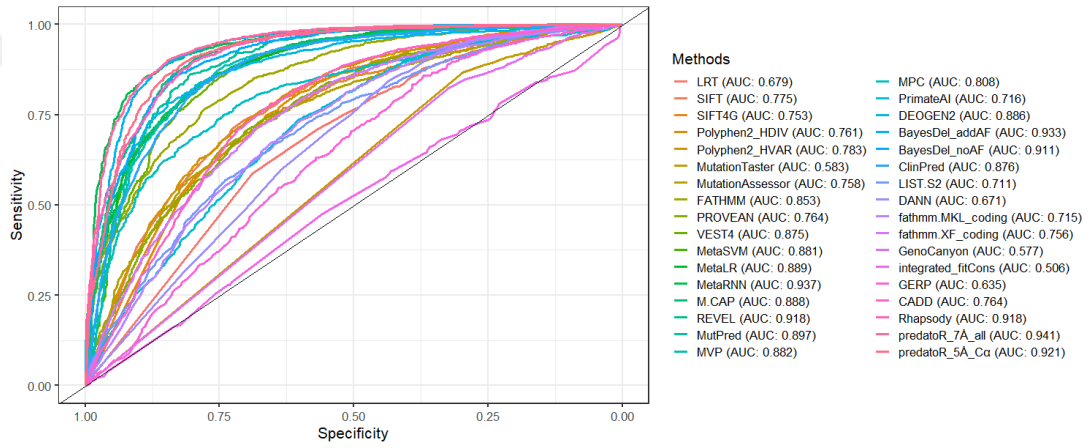

Figure 9. Comparison between 34 different impact prediction methods. Our 2 methods represented as “predatoR_7Å_all” and “predatoR_5Å_Cα”.

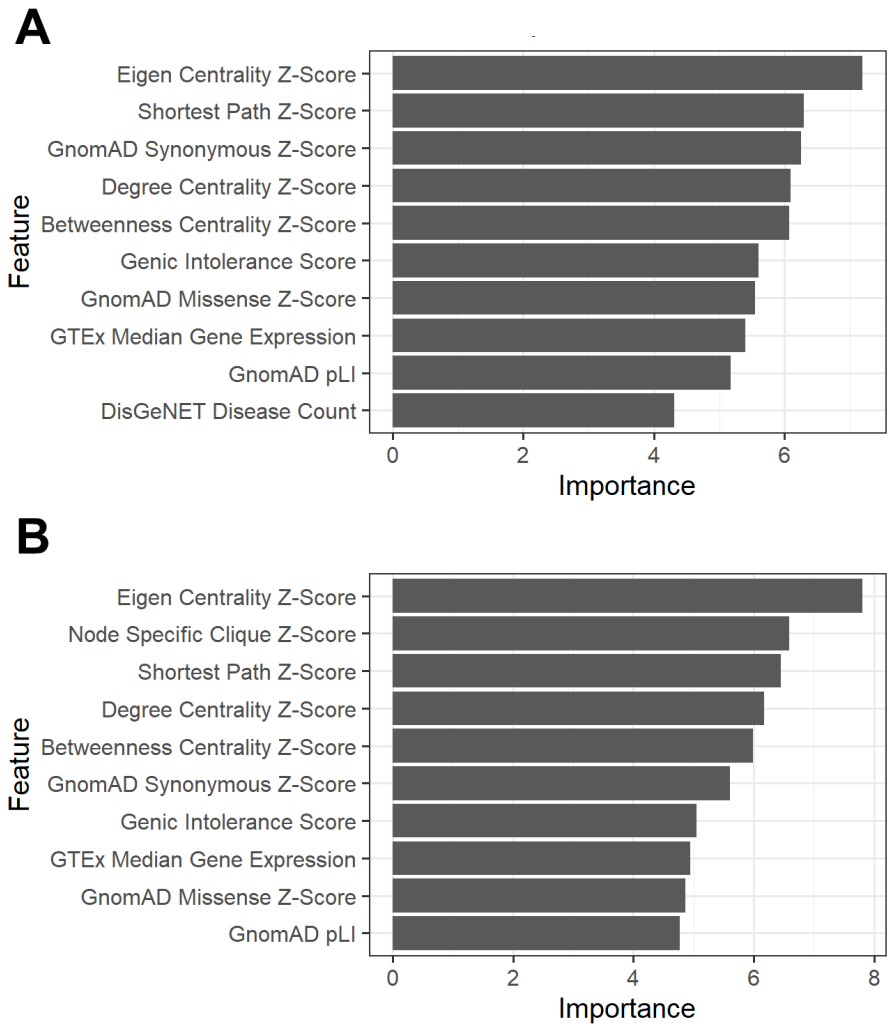

Figure 10. Feature importance plots of the 10 most important feature of the final models. (A) 7Å-all atoms approach used Adaboost model. (B) 5Å-Cα atoms approach used Adaboost model.

4.3 predatoR R Package

predatoR R package developed for making impact predictions. predatoR contains 2 models for prediction, 7Å-all atoms approach Adaboost model and 5Å- Cα atoms approach Adaboost model. User can choose which model they are going to use. predatoR package takes an input containing PDB ID, chain ID, PDB-based position, reference amino acid, mutant amino acid and gene name. Gene name information is an optional argument that the package can assign related gene name from PDB-chain

ID. If there are multiple genes associated with the same PDB-chain ID, gene having the higher gnomAD metrics across 3 metrics was selected. Simplified workflow of the predatoR package is shown in the Figure 11.

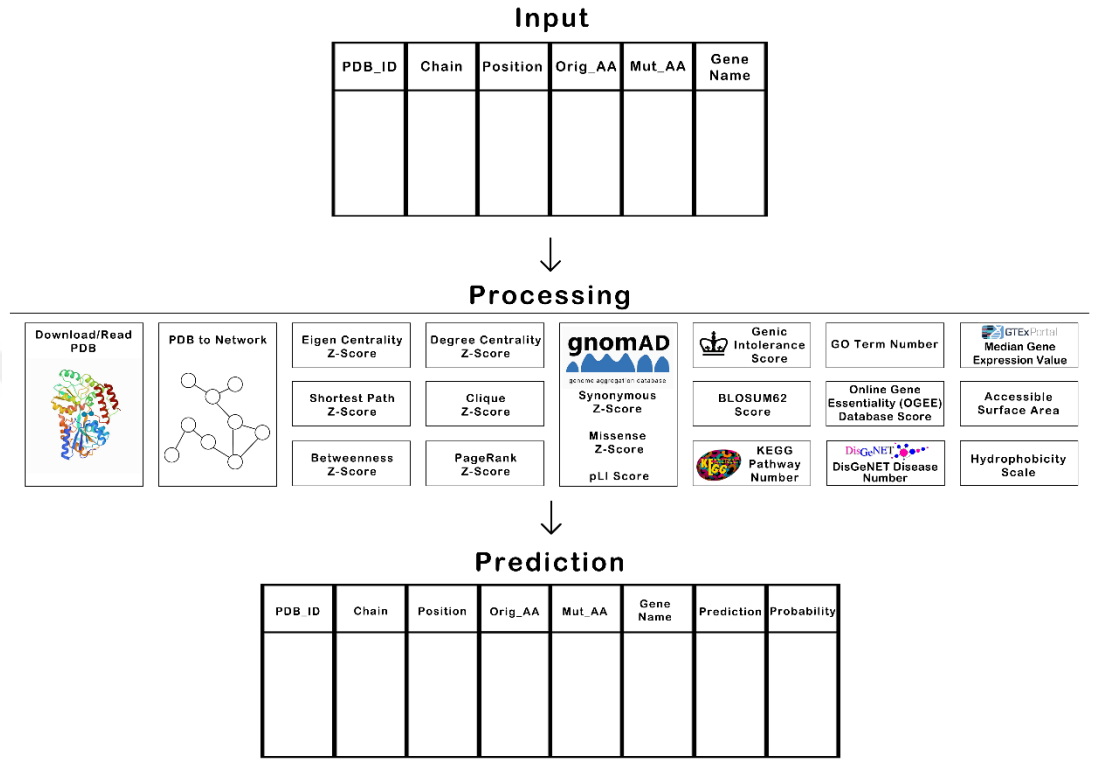

Figure 11. Simplified workflow of the package predatoR. predatoR takes an input as data frame structures and has 5 mandatory, 1 optional argument: PDB ID, chain ID, PDB-based position, reference amino acid, mutant amino acid and gene name (optional).

Package contains 19 different functions that calculates 24 features used in the model. Two functions, “read_PDB” and “PDB2connections” for reading PDB structure and calculating the distance between each atom respectively. Six functions, “betweenness_score”, “clique_score”, “degree_score” “eigen_centrality_score”, “pagerank_score” and “shortest_path_score” for building networks from the interaction list. Seven functions, “DisGeNET”, “gene_essentiality”, “genic_intolerance”, “gnomad_scores”, “GO_terms”, “GTEx” and “KEGG_pathway_number” for calculating and/or annotating gene-based features.

Two functions, “amino_acid_features” and “BLOSUM62_score” for calculating and/or annotating amino acid-based features. One function “impact_prediction” for making impact prediction and, “predatoR” function is the wrapper function for running all other functions and making predictions automatically.

“predatoR” function has 2 arguments for specifying network approach: “distance_cutoff” and “network_approach”. “network_approach” argument can be “all” or “ca”. In “all” approach, all atoms in the structure used for building network. In “ca”, atoms of the structure are filtered to contain only Cα atoms. “distance_cutoff” argument can be set any desired threshold. However, for making impact prediction, “distance_cutoff” argument need to be set as “7” or “5” and “network_approach” need to be set as “all” or “ca” respectively due to the models in the package. Moreover, package can be used for exploratory purposes. For this approach, “distance_cutoff” argument can be set any value for setting edges. Package returns a data frame contains all 24 features annotated to the input dataset. However, predictions cannot be made by using parameters except 7-all or 5-Cα.

predatoR package is available on GitHub (https://github.com/berkgurdamar/predatoR). Logo of the predatoR package is shown in the Figure 12.

Figure 12. The logo of predatoR R package.

5 DISCUSSION

Classification of a mutation is important for variant prioritization, better and faster diagnosis and understanding the mechanism behind the diseases. *In silico* methods are very important for classifying mutations due to the limitations in the experimental studies such as high cost and protein purification. Sequence-based, structure-based and thermodynamic-based approaches have been done by currently available methods and they showed the power of in silico methods.

In this thesis study, we developed a new machine learning-based method for mutation impact prediction by using network properties. Network properties calculated from protein structures by converting a PDB structure into a network. Two different interatomic interaction distance cutoffs were tested for network building, 5Å and 7Å, and 2 different approached have been made for network formalization, building networks using all atoms in the structure and Cα atoms in the structure. Best performed models were selected as final models for each network formalization approaches. Final model built with using all atoms showed better performance on the validation dataset. Also, 7Å interatomic interaction distance cutoff showed better performance than the 5Å in most of the models. This showed that building larger networks with increasing the interaction number of an atom can increase the power of predictions and performed better.

Seven different machine learning algorithms were tested for model building and testing, XGBoost, Naïve Bayes, Random Forest, Adaboost, Elastic Net, GLM and GBM. Tree-based methods, XGBoost, Random Forest, Adaboost and GBM gave higher AUROC curve values than the other remaining 3 methods. This shows that tree-based machine learning algorithms can classify mutations better with using features used in this study. Among the 4 different tree-based machine learning algorithm, Adaboost gave the best results in classifying mutations as pathogenic or neutral. Also boosting methods gave higher AUROC curve values with compared to Random Forest algorithm. It shows that with boosting algorithms better models can be created and higher accuracies and better classifications can be done.

Using network properties which were calculated from protein structures is a new and promising approach for mutation impact prediction. In this study, we accepted the protein structures as networks and with using 6 different network properties, Eigen Centrality, Betweenness Centrality, Degree Centrality, PageRank Centrality, Shortest Path Centrality and Clique Centrality, we tried to predict impact of mutations. According to the feature importance plots of the final models, among the 24 different features which were included in the final model, 4 of the 6 network properties were in the top 5 most important features for 7Å-all atoms approach used Adaboost model and 5 of the 6 network properties were in the top 5 most important features for 5Å-Cα atoms approach used Adaboost model. This shows the power of network properties on mutation impact prediction.

Eigen Centrality score was the most important feature in both final models. Eigen centrality measures the influence of a node in the network which means that effect of an atom in the structure is very important. Number of atoms which atoms connected to atom has the highest affect for classifying mutations. Shortest path represents the minimum number of steps between nodes. In this study, it represents the sum of all shortest path lengths between atoms. Reaching different atoms in a short distance have an influence on the being pathogenic or neutral variant. Degree also represents the number of interactions that a node has in the network which means that interaction number of an atom is very discriminative for impact of a mutation. Betweenness is a measure of calculating the information flow on the network. Being on more paths between atoms can also have affect the classification of a mutation.

Gene-based features was also important according to the feature importance of a 7Å-all atoms approach used Adaboost model. Two of the 3 features from gnomAD metrics were the 3rd and 7th most important features. This kind of metrics which represents the mutation acceptability have power on classification of mutations. Genic Intolerance also very similar measure like gnomAD metrics and it is also important for classification. Feature importance showed that expression of a gene and the number of diseases that a gene associated can affect the classification of a mutation.

Hydrophobicity scale and ASA values were used for calculating 6 different features. Main aim of calculating 3 different features from 1 property was to include all possible effect on the structure. In some conditions the raw value of a property can affect the pathogenicity such as ASA value of reference or mutant amino acid. On the other hand, the difference of hydrophobicity scale values between reference and the mutant amino acid can be discriminative in other conditions.

Datasets used in this study were mostly contains pathogenic variants. ClinVar dataset was used for increasing the number of neutral variants and prevent class imbalance. With an increase in the neutral variant containing datasets, better models can be built, and better methods can be developed. Moreover, different datasets contain same mutations, and those mutations were eliminated. Also, same mutation had conflicting labels (pathogenic and neutral) in different datasets. Those mutations also had to be filtered before model training. Update in the currently available databases and datasets can create positive effect on the development of new methods.

Our method requires protein structures as PDB files for making impact prediction. If the structure of a protein is not included in the PDB, structure prediction can be made by using AlphaFold (91) and predicted structures can also be used. Moreover, with using exploratory analysis approach of the predatoR, our method can be combined with using different type of features such as Thermodynamic-based information and new approaches can be made for mutation impact prediction or for any other purposes.

Our method, predatoR, outperformed 32 different mutation impact prediction methods with using their predictions on Missense3D dataset. This shows the power of network properties in classifying mutations when they were calculated from protein structures. predatoR is represented as R package for community to use. For increasing the accessibility, web server can be created for our method. Most of the mutation impact prediction methods can be used from their web servers. Web servers can decrease the complexity of the tool and increase the usage from non-R language users.

6 CONCLUSION

Impact prediction of a mutation is an important approach for classifying variations, prioritizing variants and better understanding of diseases and its mechanism. Different approaches have been made over the past decade, but their accuracies are quite low and false positive rates are high. In this study, we developed mutation impact prediction models for classifying mutations as neutral or pathogenic by using datasets collected from ClinVar and VariBench. In our approach, we converted protein structures into networks and set edges between atoms by using 2 different interatomic interaction distance cutoffs. As a result, among the 7 different machine learning methods, 2 different interatomic interaction distance cutoffs and 2 network formalization approaches, Adaboost-5Å model built with Cα atoms and Adaboost-7Å model built with all atoms showed the best performances on the test dataset and selected as final models. Final models were used for comparison with 32 different impact prediction methods. Our method outperformed currently available 32 different methods with Adaboost-7Å model built with all atoms approach. We developed an R package predatoR that makes predictions on impact of a mutation by using 24 unique features. predatoR contains 2 models, 7Å-all atoms approach used Adaboost model and 5Å-Cα atoms approach used Adaboost model. Network formalization from protein structures for mutation impact prediction is a new and promising approach. With an increase in the currently available datasets, better models and methods can be built, scientific community can get more benefit and better progress can be done in bioinformatics field.

7 REFERENCES

1. Gameiro GR, Sinkunas V, Liguori GR, Auler-Júnior JOC. Precision Medicine: Changing the way we think about healthcare. Clinics (Sao Paulo). 2018;73:e723.
2. König IR, Fuchs O, Hansen G, von Mutius E, Kopp MV. What is precision medicine? Eur Respir J. 2017 Oct;50(4):1700391.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921.
4. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013 Dec;98(6):236–8.
5. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59.
6. Akhoon N. Precision Medicine: A New Paradigm in Therapeutics. Int J Prev Med. 2021 Feb 24;12:12.
7. Reitz C. Toward precision medicine in Alzheimer's disease. Ann Transl Med. 2016 Mar;4(6):107.
8. Martiniano SL, Sagel SD, Zemanick ET. CYSTIC FIBROSIS: A MODEL SYSTEM FOR PRECISION MEDICINE. Curr Opin Pediatr. 2016 Jun;28(3):312–7.
9. Schwaederle M, Zhao M, Lee JJ, Eggermont AM, Schilsky RL, Mendelsohn J, et al. Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. J Clin Oncol. 2015 Nov 10;33(32):3817–25.
10. Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–73.
11. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.
12. Cotton RGH, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, et al. The Human Variome Project. Science. 2008 Nov 7;322(5903):861–2.
13. Tan KP, Kanitkar TR, Kwoh CK, Madhusudhan MS. Packpred: Predicting the Functional Effect of Missense Mutations. Front Mol Biosci. 2021;8:646288.
14. Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics. 2015 Sep 1;31(17):2816–21.
15. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. J Mol Biol. 2014 Jul 15;426(14):2692–701.
16. Haendel M, Vasilevsky N, Unni D, Bologa C, Harris N, Rehm H, et al. How many rare diseases are there? Nat Rev Drug Discov. 2020 Feb;19(2):77–8.
17. Danese E, Lippi G. Rare diseases: the paradox of an emerging challenge. Ann Transl Med. 2018 Sep;6(17):329.
18. Iversen ES, Couch FJ, Goldgar DE, Tavtigian SV, Monteiro ANA. A Computational Method to Classify Variants of Uncertain Significance Using Functional Assay Data With Application to BRCA1. Cancer Epidemiol Biomarkers Prev. 2011 Jun;20(6):1078–88.
19. Morales A, Hershberger RE. Variants of Uncertain Significance: Should We Revisit How They Are Evaluated and Disclosed? Circ Genom Precis Med. 2018 Jun;11(6):e002169.
20. Mahmood K, Jung C hee, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017 May 16;11:10.
21. Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics. 2016 Jun;203(2):635–47.
22. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. Human Molecular Genetics. 2001 Oct 2;10(21):2319–28.
23. Wang Z, Moult J. SNPs, protein structure, and disease. Human Mutation. 2001;17(4):263–70.
24. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol. 2001 Mar 23;307(2):683–706.

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.
26. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444–8.
27. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.
28. Wang D, Li J, Wang Y, Wang E. A comparison on predicting functional impact of genomic variants. NAR Genom Bioinform. 2022 Jan 14;4(1):lqab122.
29. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Research. 2002 Sep 1;30(17):3894–900.
30. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018 Aug;50(8):1161–70.
31. Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. 2020 May 1;36(10):3084–92.
32. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. Am J Hum Genet. 2011 Apr 8;88(4):440–9.
33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.
34. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 6;99(4):877–85.
35. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015 Aug 15;31(16):2745–7.
36. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013 Jan;34(1):57–65.
37. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Computational Biology. 2010 Dec 2;6(12):e1001025.
38. Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 2013 Oct;102(4):223–8.
39. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics. 2013 May 28;14(Suppl 3):S2.
40. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics. 2015 Apr 15;24(8):2125–37.
41. Sasidharan Nair P, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013 Jan;34(1):42–9.
42. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D980–5.
43. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000 Jan 1;28(1):352–5.
44. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013 Jan;41(Database issue):D936–41.
45. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J Mol Biol. 2019 May 17;431(11):2197–212.
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.
47. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019 Dec 21;19:281.
48. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.

49. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):160.
50. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.
51. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. San Francisco California USA: ACM; 2016 [cited 2022 Jun 6]. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785
52. Freund Y, Schapire RE. A Short Introduction to Boosting. :14.
53. Cao Y, Miao QG, Liu JC, Gao L. Advance and Prospects of AdaBoost Algorithm. Acta Automatica Sinica. 2013 Jun;39(6):745–58.
54. Freund Y, Schapire RE. Experiments with a New Boosting Algorithm. 1996.
55. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007 Aug 15;73(16):5261–7.
56. Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society Series A (General). 1972;135(3):370.
57. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001 Oct;29(5):1189–232.
58. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013 Dec 4;7:21.
59. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B. 2005 Apr;67(2):301–20.
60. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet. 2013 Dec 4;4:270.
61. Batool K, Niazi MA. Towards a Methodology for Validation of Centrality Measures in Complex Networks. Perc M, editor. PLoS ONE. 2014 Apr 7;9(4):e90283.
62. Bloch F, Jackson MO, Tebaldi P. Centrality Measures in Networks [Internet]. Rochester, NY: Social Science Research Network; 2019 Jun [cited 2022 May 5]. Report No.: 2749124. Available from: https://papers.ssrn.com/abstract=2749124
63. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. Briefings in Bioinformatics. 2021 Mar 22;22(2):1639–55.
64. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 1998 Apr;30(1–7):107–17.
65. Bianchini M, Gori M, Scarselli F. Inside PageRank. ACM Trans Internet Technol. 2005 Feb;5(1):92–128.
66. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, et al. Network-based prediction of protein interactions. Nat Commun. 2019 Mar 18;10:1240.
67. Aydınkal RM, Serçinoğlu O, Ozbek P. ProSNEx: a web-based application for exploration and analysis of protein structures using network formalism. Nucleic Acids Research. 2019 Jul 2;47(W1):W471–6.
68. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/
69. RStudio Team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: RStudio, PBC; 2021. Available from: http://www.rstudio.com/
70. Stephenson JD, Laskowski RA, Nightingale A, Hurles ME, Thornton JM. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. Bioinformatics. 2019 Nov 1;35(22):4854–6.
71. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. BMC Genomics. 2009 Jan 14;10(1):22.
72. Martin ACR. Mapping PDB chains to UniProtKB entries. Bioinformatics. 2005 Dec 1;21(23):4297–301.
73. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006 Nov 1;22(21):2695–6.
74. Csardi G, Nepusz T. The igraph software package for complex network research. :9.
75. The Genome Aggregation Database (gnomAD) [Internet]. [cited 2022 May 2]. Available from: https://www.nature.com/immersive/d42859-020-00002-x/index.html

76. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017 Jan 4;45(D1):D353–61.
77. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
78. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017 Jan 4;45(D1):D833–9.
79. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLOS Genetics. 2013 Aug 22;9(8):e1003709.
80. Gurumayum S, Jiang P, Hao X, Campos TL, Young ND, Korhonen PK, et al. OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. Nucleic Acids Research. 2021 Jan 8;49(D1):D998–1003.
81. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.
82. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Human Mutation [Internet]. [cited 2022 May 5];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.24309
83. Geistlinger L, Csaba G, Zimmer R. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinformatics. 2016 Jan 20;17(1):45.
84. Juretić D, Lučić B, Zucić D, Trinajstić N. Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. In: Theoretical and Computational Chemistry [Internet]. Elsevier; 1998 [cited 2022 Jul 28]. p. 405–45. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1380732398800150
85. Radzicka A, Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry. 1988 Mar 8;27(5):1664–70.
86. D.M RSR. aaSEA: Amino Acid Substitution Effect Analyser [Internet]. 2019 [cited 2022 May 2]. Available from: https://CRAN.R-project.org/package=aaSEA
87. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915–9.
88. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28(5):1–26.
89. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.
90. Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, et al. TransVar: a multilevel variant annotator for precision genomics. Nat Methods. 2015 Nov;12(11):1002–3.
91. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

8 CURRICULUM VITAE