Abstract ID:- 71

Abstract Topic:- Cancer

Abstract Title:- Classification of Pathogenicity from Germline Missense Variants using Machine Learning Algorithms

Presenting author name :- Brindha Senthil Kumar

Presenting author institute:- Department of Computer Engineering, Mizoram University

Aims:-The present study had proposed a novel machine learning-based tool to classify between benign and pathogenic missense variants.

Methods:- There were 222,540 germline missense variants obtained from 64 gastric cancer patients via next generation sequencing (NGS). The missense variants were subjected to pathogenicity prediction tools: MutationTaster, FATHMM, LR, LRT, RadialSVM, SIFT, and Polyphen2 to acquire appropriate class labels (pathogenic or benign). A missense variant was labeled pathogenic if and only if all the above-seven pathogenicity prediction tools predicted it as pathogenic; otherwise it had been labeled as benign. The dataset contained 200171 benign and 22369 pathogenic missense variants with 62 features. The dataset was randomly divided into 70% for training and 30% for testing. Extra trees classifier algorithm was implemented to extract the features of importance from 62 independent features using 5-fold cross validation via gridsearchCV on training dataset and obtained the best estimators. A set of 5-ensemble algorithms was chosen to classify between pathogenic and benign missense variants: Random Forest (RF), Bagging classifier (BC), Extra Trees (ET), AdaBoost (AB) and Gradient Boosting (GB). The dataset was divided into five sets based on feature significance: top 30 features, top 20 features, top 10 features, top 8 features, and all features.

Results:- The five ensemble models' hyper-parameters were tuned using 5-fold gridsearhCV on training dataset: ExtraTreesClassifier(max_depth=100, n_estimators=50), BaggingClassifier(estimator=DecisionTreeClassifier(), n_estimator=100), RandomForestClassifier(max_depth=50, n_estimators=50), AdaBoostClassifier(n_estimators=1000) and GradientBoostingClassifier(n_estimators=1000). RF, BC and ET models had shown outstanding performances on all five test datasets: accuracy, precision, recall and f1_score of 99% each individual evaluation metric with Matthew's correlation coefficient (MCC) of 1.0 and precision_recall (PR) curve of 1.0 for both benign and pathogenic classes. The presented models showed that top 8 features: phyloP46way_placental, VEST3_score, ExAC_AFR, AFR.sites.2015_08, AF_afr, SiPhy_29way_logOdds, and ALL.sites.2015_08 scores can be playing a significant role in determining the pathogenicity of the missense variant.

Conclusions:- The performance of generated in-house machine learning tool had clearly paved way to improve the patients' diagnosis rate and to identify novel disease-specific variants with high probability; furthermore, this knowledge transfer will aid in personalized-precision genomic medicine to reduce the cancer incidence and early diagnosis.

Keywords:- germline missense variants, gastric cancer, machine learning, pathogenicity, precision_recall curve