Homology-based prediction of resistance to antituberculous medications using machine learning algorithms

Mohamed Dafaalla (✉ mdafaallah200@gmail.com)
https://orcid.org/0000-0003-4625-2775

Mohamed Omer Elsadiq Abdullah
Institute of Endemic Diseases

Sahar Bakhiet
Institute of Endemic Diseases

Muntasir Ibrahim
Institute of Endemic Diseases

Abstract

Objectives: We aimed to develop a prediction model based on machine learning algortihms to predict the impact of variants on resistance of Mycobacterium. Data was collected from TB Drug Resistance Database (TBDReaMDB), and the drug sensitive variants from GMTV database. We also collected a List of 1115 NsSNPS reported in proteins related to drug resistance to Rifampicin, Isoniazid, Pyrazinamide and Ethambutol. PMUT online tool was used to generate the features included in the algorithm training. We trained different classifiers using R software on the features generated by PMUT. The classifiers trained are Random Forrest, Boosting prediction, Naive Bayes, Neural networks, k-Nearest Neighbors, Logistic regression, and Linear Discriminant analysis.

Results: The 445 variants valid for comparison were divided into training dataset (75%) and testing dataset (25%). We compared the classifiers according to the AUC, accuracy, kappa, sensitivity, specificity, positive predictive value, and negative predictive value. Results show that random forrest is the best classifier (accuracy: 0.9072. Kappa: 0.690, Sensitivity: 1.00, Specificity : 0.5909, Pos-Pred Value : 0.8929, Neg-Pred Value : 1.00, Detection Rate : 0.773). This indicates that Homology-based machine learning algorithms could be a solid base for development of rapid tools for screening of M.TB resistance to medications.

Introduction

The most common cause of death attributed to a single microorganism worldwide is Tuberculosis (TB). Without combatting antimicrobial resistance in Mycobacterium tuberculosis the goals of the End TB strategy of the World Health Organization (WHO) will not be achieved. The antimicrobial resistance was acknowledged as an important threat to public health and economic growth by the leaders of the G20 nations(1). The WHO estimates that of the 600 000 newly emerging MDR-TB or rifampicin resistant TB (RR-TB) cases every year, only 25% are detected and notified. Evidence also suggests that there is a 16% and 80% higher frequency among estimated and notified MDR-TB patients. In addition, the number of reported MDR/RR-TB peaked by more than one third in 9 of the 30 high MDR-TB burden countries in the period between 2015 and 2016(1).

The definite diagnosis of pulmonary tuberculosis is by isolation of _M. tuberculosis_ from a bodily secretion (e.g., culture of sputum, bronchoalveolar lavage, or pleural fluid) or tissue (pleural biopsy or lung biopsy). Additional diagnostic tools are considered sufficient to make a diagnosis. These include sputum acid-fast bacilli (AFB) smear and nucleic acid amplification (NAA) testing. Radiographic studies are also important supportive tool(2).

The implementation of molecular tests worldwide has dramatically reduced the time needed to diagnose tuberculosis(3−6). These tests can be used to diagnose the infection in the different biological samples (sputum, cerebrospinal fluid, pleural fluid, urine, blood, stool)(6). Nucleic acid amplification tests (NAATs) also enable the prediction of resistance to important first- and second-line drugs on the basis of resistance-conferring mutations(7). Nevertheless, most commercially available tests are available only to certain drug targets and remain inadequate when comprehensive individualized treatment regimens have to be established(1).

The best coverage for molecular DST is given by sequencing of the entire mycobacterial genome (whole genome sequencing, WGS)(8). This method reports all potential resistance-conferring mutations. However, there are few drawbacks for the WGS. First, WGS requires relatively large amounts of enriched mycobacterial DNA, which indicates early positive culture materials, to sequence. Second, although special enrichment procedures make direct sequencing from fresh clinical samples applicable, it is still several years away from clinical implementation(9−11). Researchers believe that If these issues can be overcome, information obtained from WGS with computational sequence analysis, which may be uploaded to large databases, is probably the future of molecular drug resistance prediction(12,13).

The aim of this study was to develop a bioinformatics machine learning algorithm that is capable of prediction of variants impact on resistance to antituberculous medications. The next sections illustrate the methods used and the main features included in the classifier beside the accuracy parameters.

Methods

Study design

I developed a machine learning algorithm that is able to predict the impact of Nonsynonymous (NsSNPS) variants on susceptibility of Mycpbacterium Tuberculosis to antituberculous medications based on amino acid sequence and predicted protein secondary structure.

Database: -

We collected data from 2 databases; from TB Drug Resistance Database (TBDReaMDB), and collected the drug sensitive variants from GMTV database(14,15).

TBDReaMDB is a comprehensive resource on drug resistance mutations in *M. tuberculosis* developed by conducting a systematic review to identify drug resistance mutations from the existing literature to include in the database(14).

GMTV contains a broad spectrum of data derived from different sources and related to *M. tuberculosis* molecular biology, epidemiology, TB clinical outcome, year and place of isolation, drug resistance profiles and displays the variants across the genome using a dedicated genome browser. GMTVdatabase, which includes 1084 genomes and over 69,000 SNP or Indel variants, can be queried about *M. tuberculosis* genome variation and putative associations with drug resistance, geographical origin, and clinical stages and outcomes (15).

Inclusion and Exclusion criteria:

We collected a List of 1488 NsSNPS that are associated with drug resistance to Rifampicin, Isoniazid, Pyrazinamide and Ethambutol. The data included the gene ID, protein ID, codon number, wild amino acid, mutation amino acid, drug susceptibility, and variant impact on protein.

Variants were grouped into two groups (sensitive vs resistant) according to drug sensitivity. Variants found in drug sensitive organism were labelled sensitive, whereas variants found in drug sensitive organism were labelled resistant. We included only proteins that have variants in the two groups. These proteins are Rv0667, Rv1908c, Rv2043c, Rv2428, Rv3793, and Rv3795. The final number of variants included was 1115.

Features generation:

We used PMUT online tool to generate the features included in the algorithm training.

PMut Web portal allows the user to perform pathology predictions, to access a complete repository of pre-calculated predictions, and to generate and validate new predictors. The default predictor performs with good quality scores. The PMut portal is freely accessible at http://mmb.irbbarcelona.org/PMut (16).

The features computed by PMUT and entered in the classifiers were number of sequences in the alignment, number of amino acids in the aligned position (no gaps), total and relative number of aligned wild type amino acids, total and relative number of aligned mutated amino acids, position Weight Matrix score, and PMUT overall score. Table (1) below shows the selected features for the PMut2017 predictor.

Classifiers training and testing

We trained different classifiers on the features generated by PMUT and compared them according to the confusion matrix, receiver operator curve (ROC), and area under the curve (AUC). The classifiers that were trained are Random Forrest (rf), Boosting prediction(ada), Naive Bayes (nb), Neural networks(nnet), k-Nearest Neighbors (knn), Logistic regression (LR), and Linear Discriminant analysis (lda).

Software/ Packages:

We used Microsoft excel 2016 and SPSS v22 to clean the data. We used R Software to train the classifiers. We used the "caret" and "nnet" packages to train and evaluate the classifiers, and "ROCR" package to plot the ROCs curves and calculate the AUC for each classifier.

Results
Exploratory analysis:

Exploratory data analysis revealed that all (791 mutation) frameshift mutations, insertions, and deletions are associated with drug resistance and none was reported in drug sensitive organism. Regarding STOP variants, only 5 variants were found in drug sensitive organism. As a result; they were not included in the training and testing datasets. We divided the remaining 445 variants into two datasets: training dataset (75%) and testing dataset (25%). Table (2) shows proteins of variants included in the training and testing datasets

Classifiers evaluation:

After training the classifiers on the training dataset we assessed the classifiers prediction on the test dataset. We compared the classifiers according to the AUC, accuracy, kappa, sensitivity, specificity, positive predictive value, and negative predictive value. Table (3) below compares the prediction accuracy of the classifiers. It shows that the Random Forrest are superior to other classifiers. Table (3) and figure S1 compares the performance of the classifiers. Figure S1 which shows the ROC of all classifiers is attached as a supplementary file.

Discussion

A machine learning based algorithm that predict the impact of variants on the sensitivity to the antituberculous medications was developed and validated. This could be the infrastructure for a bioinformatics tool that can reduce the time and cost of diagnosing drug resistant mycobacteria.

Construction of this algorithm on the concept of conservation of amino acids is one of the strong features of this algorithm. This is because antituberculous medications have to interact directly with mycobacterial proteins. They either act by binding a protein directly, need a protein to be activated. For example, Rifampin is the cornerstone of short-course chemotherapy regimens, so rifampin resistance prolongs and complicates treatment. Rifampin is thought to act against M. tuberculosis by binding to RNA polymerase, leading to interference with transcription and RNA prolongation (12,17). Similarly, Ethambutol acts by inhibiting the Arabinosyl transferase. Arabinosyl transferase is an enzyme that polymerizes arabinose into arabinan and then arabinogalactan, a mycobacterial cell wall constituent (18–20). Both Isoniazide and Pyrazinamide requires modifications by catalase and pyrazinamidase respectively(21–23).

The sensitivity of the direct microscopy and of the mycobacterium culture for a bronchoalveolar lavage sample is 19% and 50% respectively while the specificity is 90% and 100% respectively which make both of them excellent diagnostic tools but not suitable for tuberculosis screening(24). The features of our algorithms are exactly the opposite. The algorithm was accurately able to predict the impact of the variant on the drug resistance. It was able to identify all the drug resistant variants with a sensitivity of 100%. However, the specificity was just around sixty percent. This indicates that there is considerable possibility of classifying a drug sensitive variant as drug resistant, whereas the probability of classifying a drug resistant variant as a sensitive variant is extremely low. These features leads to a conclusion that machine learning based classifiers can be an excellent base for screening tools for mycobacterium drug resistance.

The question that rise up to the surface is what could be the next step. The first step will be developing an online server that can be used for prediction of the variants impact on the drug resistance instantly within minutes. If this server is combined with a molecular tool that can detect the aminoacids variants we will end up with a bedside quick screening tool for tuberculosis drug resistant.

- Limitations:

The number of variants were finally incorporated in the classifier training was relatively smaller compared to previously validated bioinformatics tools(25–37). This also affected the number of proteins included. This can be updated in the future when additional information about variants with known impact on drug resistance are available.

Declarations

Ethical approval and consent to participate:-

Not applicable as no human subjects were involved and data was extracted from open access database.

Availability of data and material:

Datasets, figure (1), and R script are available as supplementary material.

Funding:

No funds to disclose

Acknowledgment:

Sincere appreciation to Institute of Endemic Diseases staff.

Consent for Publication:-

The corresponding author has the consent for submitting the manuscript for publication

Competing interests:

No interests to disclose.

Author's contribution:

All authors were involved in idea formation, manuscript writing and revision.

References

1. Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29641838

2. Sharma A, Hill A, Kurbatova E, van der Walt M, Kvasnovsky C, Tupasi TE, et al. Estimating the future burden of multidrug-resistant and extensively drug-resistant tuberculosis in India, the Philippines, Russia, and South Africa: a mathematical modelling study. Lancet Infect Dis. 2017 Jul;17(7):707–15.

3. Boehme CC, Nicol MP, Nabeta P, Michael JS, Gotuzzo E, Tahirli R, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. Lancet Lond Engl. 2011 Apr 30;377(9776):1495–505.

4. Hillemann D, Rüsch-Gerdes S, Boehme C, Richter E. Rapid molecular detection of extrapulmonary tuberculosis by the automated GeneXpert MTB/RIF system. J Clin Microbiol. 2011 Apr;49(4):1202–5.

5. Rapid molecular detection of tuberculosis and rifampin resistance. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20825313

6. Steingart KR, Sohn H, Schiller I, Kloda LA, Boehme CC, Pai M, et al. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database Syst Rev. 2013 Jan 31;(1):CD009593.

7. Clinical implications of molecular drug resistance testing for Mycobacterium tuberculosis: a TBNET/RESIST-TB consensus statement. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26688526

8. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26669893

9. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25631807

10. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol. 2017;55(5):1285–98.

11. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol. 2015 Jul;53(7):2230–7.

12. Donnabella V, Martiniuk F, Kinney D, Bacerdo M, Bonk S, Hanna B, et al. Isolation of the gene for the beta subunit of RNA polymerase from rifampicin-resistant Mycobacterium tuberculosis and identification of new mutations. Am J Respir Cell Mol Biol. 1994 Dec;11(6):639–43.

13. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25854485

14. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis Drug Resistance Mutation Database. PLOS Med. 2009 Feb 10;6(2):e1000002.

15. Chernyaeva EN, Shulgina MV, Rotkevich MS, Dobrynin PV, Simonov SA, Shitikov EA, et al. Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics. 2014 Apr 25;15:308.

16. López-Ferrando V, Gazzo A, de la Cruz X, Orozco M, Gelpí JL. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. Nucleic Acids Res. 2017 Jul 3;45(W1):W222–8.

17. Zaw MT, Emran NA, Lin Z. Mutations inside rifampicin-resistance determining region of rpoB gene associated with rifampicin-resistance in Mycobacterium tuberculosis. J Infect Public Health. 2018 Apr 26;

18. Ramaswamy S, Musser JM. Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update. Tuber Lung Dis Off J Int Union Tuberc Lung Dis. 1998;79(1):3–29.

19. Belanger AE, Besra GS, Ford ME, Mikusová K, Belisle JT, Brennan PJ, et al. The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc Natl Acad Sci U S A. 1996 Oct 15;93(21):11919–24.

20. The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term = The+emb+operon%2C+a+gene+cluster+of+Mycobacterium+tuberculosis+involved+in+resistance+to+ethambutol.

21. Rozwarski DA, Grant GA, Barton DH, Jacobs WR, Sacchettini JC. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. Science. 1998 Jan 2;279(5347):98–102.

22. Scorpio A, Zhang Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med. 1996 Jun;2(6):662–7.

23. Scorpio A, Lindholm-Levy P, Heifets L, Gilman R, Siddiqi S, Cynamon M, et al. Characterization of pncA mutations in pyrazinamide-resistant Mycobacterium tuberculosis. Antimicrob Agents Chemother. 1997 Mar;41(3):540–3.

24. Jacomelli M, Silva PRAA, Rodrigues AJ, Demarzo SE, Seicento M, Figueiredo VR. Bronchoscopy for the diagnosis of pulmonary tuberculosis in patients with negative sputum smear microscopy results. J Bras Pneumol. 2012 Apr;38(2):167–73.

25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

26. Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A. 2004 Oct 26;101(43):15398–403.

27. Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term = Development+and+Validation+of+a+Computational+Method+for+Assessment+of+Missense+Variants+in+Hypertrophic+Cardiomyopathy

28. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term = Yuan+HY%2C+Chiou+JJ%2C+Tseng+WH%2C+et+al.+FASTSNP%3A+an+always+up-to-date+and+extendable+service+for+SNP+function+analysis+and+prioritization.+Nucleic+Acids+Res+2006%3B34%3A

29. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002 Sep 1;30(17):3894–900.

30. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W480–482.

31. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17130144

32. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005 Jul;15(7):978–86.

33. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinforma Oxf Engl. 2006 Nov 15;22(22):2729–34.

34. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comput Biol [Internet]. 2014 Jan 16 [cited 2018 Jul 16];10(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894168/

35. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

36. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823–35.

37. SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. - PubMed - NCBI [Internet]. [cited 2018 Jul 16]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term = SNPeffect+v2.0%3A+a+new+step+in+investigating+the+molecular+phenotypic+effects+of+human+non-synonymous+SNPs.

Tables

Table (1): Selected features for the PMut2017 predictor.

#	Alignment	Database	Filter	Value	Weighted
1	PSI-BLAST	UniRef100	E-value < 10^{-75}	Number of amino acids in the aligned position.	No
2	PSI-BLAST	UniRef100	E-value < 10^{-75}	Position Weight Matrix score.	No
3	PSI-BLAST	UniRef100	E-value < 10^{-75}	Proportion of wild type amino acids in the aligned position.	BLAST score
4	PSI-BLAST	UniRef90	E-value < 10^{-45}	Number of amino acids in the aligned position.	No
5	MSA (Kalign2)	UniRef100	All	Number of sequences in the alignment.	No
6	MSA (Kalign2)	UniRef100	All	Number of wild type amino acids in the aligned position.	No
7	MSA (Kalign2)	UniRef100	All	Position Weight Matrix score.	Sequence similarity
8	MSA (Kalign2)	UniRef100	Human	Number of amino acids in the aligned position.	No
9	MSA (Kalign2)	UniRef90	All	Number of amino acids in the aligned position.	Sequence similarity
10	MSA (Kalign2)	UniRef90	All	Number of wild type amino acids in the aligned position.	Sequence similarity
11	MSA (Kalign2)	UniRef90	Human	Number of amino acids in the aligned position.	Sequence similarity
12	Miyata substitution matrix score.				

Table (2): Proteins of variants included in the training and testing datasets

Gene name	Protein name	No of SNPs
embB	Arabinosyltransferase B	132
embC	Arabinosyltransferase C	33
katG	Catalase-peroxidase	221
rpoB	RNA polymerase	59
Total		445

Table (3): Performance evaluation of the classifiers

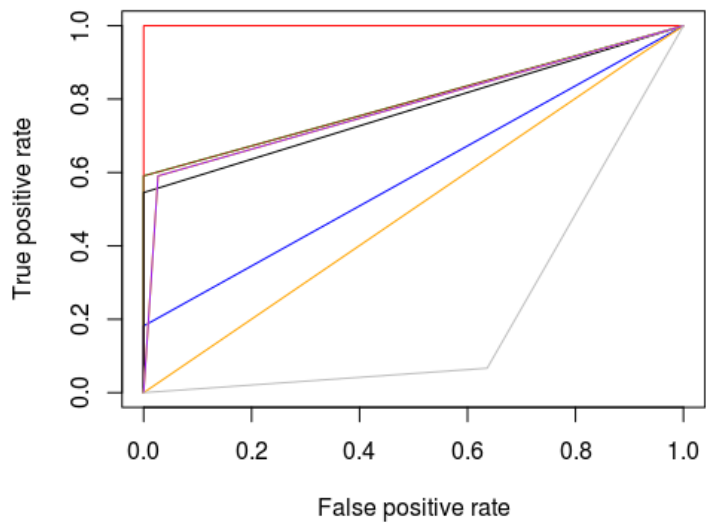|	Boosting (ada)	K nearst neighbor (knn)	Linear discriminant analysis (lda)	Logistic regression (LR)	Naïve Bayes (nb)	Neural networks (nnet)	Random Forrest (rf)	Support vector machine (svm)
Accuracy	0.9072	0.8969	0.86	0.2784	0.8144	0.9072	0.9072	0.8866
95% CI	(0.8312, 0.9567)	(0.8186, 0.9494)	(0.7817, 0.9267)	(0.1921, 0.3786)	(0.7227, 0.8862)	(0.8312, 0.9567)	(0.8312, 0.9567)	(0.8061, 0.942)
Kappa	0.6908	0.6498	0.598	0.0012	0.2558	0.6908	0.690	0.6357
Mcnemar's Test P-Value	0.0076608	0.004427	0.57	7.912e-15	6.151e-05	0.0076608	0.0076608	0.070440
Sensitivity	1.00	1.0	0.93	0.09333	1.0000	1.0000	1.0000	0.9733
Specificity	0.5909	0.54	0.63	0.90909	0.1818	0.5909	0.5909	0.5909
Pos Pred Value	0.8929	0.88	0.897	0.77778	.8065	0.8929	0.8929	0.8902
Neg Pred Value	1.0000	1.0	0.73	0.227	1.0000	1.0000	1.0000	0.8667
Detection Rate	0.7732	0.77	0.721	0.07216	0.7732	0.773	0.7732	0.7526
AUC	0.7954545	0.7727273	0.2151515	0.6742424	0.5909091	0.7954545	0.8330303	0.7821212

Figures

Figure 1

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- trainingdataset.csv
- testingdataset.csv