Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data

by Wiwien Hadikurniawati

Submission date: 03-Sep-2022 06:24PM (UTC+0700)
Submission ID: 1891836352
File name: Hadikurniawati_2021_J._Phys.__Conf._Ser._1869_012093.pdf (684.26K)
Word count: 2063
Character count: 10475

Journal of Physics: Conference Series

PAPER · OPEN ACCESS

Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data

To cite this article: W Hadikurniawati *et al* 2021 *J. Phys.: Conf. Ser.* **1869** 012093

Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data

W Hadikurniawati[1,*], M T Anwar[1], D Marlina[2] and H Kusumo[3]

[1] Faculty of Information Technology, Universitas Stikubank, Jl. Tri Lomba Juang No 1 Semarang, Indonesia
[2] Faculty of Pharmacy, Universitas Setia Budi, Jl. Letjen Sutoyo, Mojosongo, Kec. Jebres, Kota Surakarta, Indonesia
[3] Department of Informatics Management, Universitas Stekom, Jl. Majapahit 605, Kec. Pedurungan, Semarang, Indonesia

*wiwien@edu.unisbank.ac.id

Abstract. Tuberculosis is a serious infectious disease caused by Mycobacterium tuberculosis (MTB) that primarily affects the lungs. It is known that several strains of MTB are resistant to drugs used in the treatment. This situation calls for the importance to detect and prevent further drug resistance and thus reducing the mortality rate. The conventional molecular diagnostic test is costly, requires a long time to conduct, and has low prediction ability. This research aims to explore the Machine Learning approach to accurately predict drug resistance which offers a much faster and cheaper solution than the conventional one. Experiments were carried out on 3393 isolates of MTB using several Machine Learning algorithms including C4.5, Random Forest, and Logitboost. Multiple drugs evaluated in this study include rifampicin (RIF), isoniazid (INH), pyrazinamide (PZA), and ethambutol (EMB). By using 10-fold cross-validation, the result had demonstrated that the model can accurately predict drug resistance with an accuracy of 99% and with Area Under Curve (AUC) reaching (near) 1. This result suggests that Machine Learning approach has a promising result in predicting Tuberculosis drug resistance.

1. Introduction

Tuberculosis is a serious infectious disease caused by Mycobacterium tuberculosis (MTB) that primarily affects the lungs and is one of the most deadly infectious disease in the world [1]. It is known that several strains of MTB are resistant to drugs used in the treatment [2]. This situation calls for the importance to detect and prevent further drug resistance and thus reducing the mortality rate. The conventional molecular diagnostic test is costly, requires a long time to conduct, and has low prediction ability. The whole-genome sequencing (WGS) captures the known and rare mutation of the MTB isolates that may contribute to the drug resistance. These mutations are used as the features for classifying the isolates if they are resistant to a drug. This research aims to explore Machine Learning (ML) techniques to accurately predict drug resistance which offers a much faster and cheaper solution than the conventional techniques.

2. Methods

Genetic data of 3393 MTB isolates were retrieved from Kaggle. Multiple drugs evaluated in this study include the first-line drugs, i.e. rifampicin (RIF), isoniazid (INH), pyrazinamide (PZA), and ethambutol

(EMB). Positive classes for RIF, INF, and PZA are 61%, 54%, 66%, and 71% respectively. The data have 222 columns representing the mutation sites alongside with the resistance class for each of the drugs. The original data is coded with [0,1] but for the processing with Waikato Environment for Knowledge Analysis (WEKA) software, we convert it to [F, T] respectively. The sample of the data is shown in Table 1. Entries with missing values are omitted or not omitted depending on whether the ML technique can handle missing values. Three ML methods are used namely C4.5, Random Forest (RF), and Logitboost. C4.5 is a classification method based on tree structure introduced by Ross Quinlan [3]. Recent research had used C4.5 for wildfire modeling [4] and rain modeling [5]. RF is an ensemble classification model that uses multiple trees to predict classes and use votes from those trees to determine the final class label. Logitboost is a boosting technique that uses decision stumps (decision tree with a single internal node). It is introduced by Friedman et.al. in 2020 [6]. The experiments were carried out by using the WEKA software [7] and the scikit-learn Machine Learning library in Python [8]. The performance evaluations of the model are done by using metrics such as Precision, Accuracy, and Area Under Curve (AUC).

Table 1. Sample of the data.

mutation1	mutation2	…	mutation222	RIF
F	T	…	F	T
T	F	…	T	F
F	F	…	F	T

3. Results and discussion

Table 2 shows the performance comparison of the models on each drug using the 10-fold cross-validation technique. This result showed that the best model performance is specific to the data, although the difference is minuscule. This result disagrees with previous research that certain models are better than the other, e.g. ensemble vs single tree [9], Random Forest [10], Logistic regression, and gradient tree boosting [11], although not yet tested against WDNN which performed better than regularized logistic regression and random forest [12]. This study concluded that model performance is data-specific which is also stated by Hicks et al [13]. However, this research produced a better result than recent research [14]. The best methods in this research produced an average of 0.975 AUC on the first line drugs which only slightly lower than other research where Logistic Regression and MD-WDNN performed best with an average AUC of 0.979 [15]. Figure 1 shows a comparison of the best and worst model's AUC.

Next, additional parameter tunings were done using the scikit-learn Machine Learning library in Python. These experiments were done using a test split of 0.1. When tuning the parameter for RF with n_trees = 74 (222/3) and n_tress = 50, the best result can have the AUC up to 1. The RF model accuracy on different n_trees are shown in Table 3. This result again showed that the model performance is data-specific (although can be minor) and can be affected by the parameter setting as also mentioned in research on Random Forest [16]. It is concluded that parameter tuning can produce (slightly) better model performance.

Table 2. Performance comparison of the models on each drug.

Drugs	Model	Precision	Accuracy (%)	AUC
RIF	J48	**0.967**	**96.67**	0.972
	J48 (min 10 cases)	0.956	95.60	0.977
	RF	0.951	95.00	**0.99**
	Logitboost	0.950	95.00	**0.99**
INH	J48	0.960	95.95	0.967
	J48 (min 10 cases)	**0.961**	**96.10**	0.961
	RF	0.951	95.10	**0.985**
	Logitboost	**0.961**	**96.10**	0.982

Table 2. Cont.

Drugs	Model	Precision	Accuracy (%)	AUC
PZA	J48	0.923	92.21	0.923
	J48 (min 10 cases)	**0.925**	**92.49**	0.943
	RF	0.919	91.74	**0.959**
	Logitboost	0.908	90.82	0.944
EMB	J48	0,916	91.44	0.916
	J48 (min 10 cases)	0,918	91.62	0.945
	RF	**0,922**	**91.96**	**0.967**
	Logitboost	0,911	91.20	0.963

Table 3. RF model accuracy on different n_trees.

Drug	n_trees = 10		n_trees = 15		n_trees = 74		n_trees = 50	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
RIF	0.970	0.99	0.970	0.99	**0.985**	**1.00**	**0.985**	**1.00**
INH	0.945	0.97	0.955	**0.98**	**0.960**	0.97	**0.960**	0.97
PZA	0.964	**0.95**	0.964	0.94	**0.974**	0.94	**0.974**	0.93
EMB	**0.985**	**0.99**	0.979	**0.99**	**0.985**	**0.99**	**0.985**	**0.99**

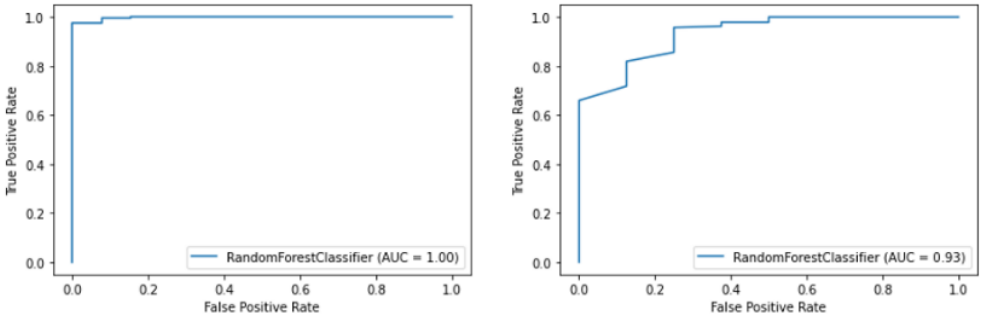

Figure 1. Comparison of the best and worst model's AUC.

4. Conclusion

We experiment on using ML techniques to predict MTB drug resistance based on DNA data. The result had demonstrated that ML techniques can accurately predict drug resistance with an accuracy of up to 99% and with Area Under Curve (AUC) reaching (near) 1. This result suggests that Machine Learning approach has a promising result in predicting Tuberculosis drug resistance. The result also showed the model performance is data-specific and that parameter tuning can result in a (slightly) better performance.

References

[1] WHO 2017 Global tuberculosis report 2018 *Glob. Tuberc. Rep. 2017*
[2] Gygli S M, Borrell S, Trauner A and Gagneux S 2017 Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives *FEMS Microbiol. Rev.* **41** 354–373
[3] Quinlan J R 2014 *C4. 5: programs for machine learning* (Elsevier)
[4] Anwar M T, Pumomo H D, Prasetyo S Y J and Hartomo K D 2018 Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*

(IEEE) pp 409–415
[5] Anwar M T, Nugrohadi S, Tantriyati V and Windarni V A 2020 Rain Prediction Using Rule-Based Machine Learning Approach *Adv. Sustain. Sci. Eng. Technol.* **2**
[6] Friedman J, Hastie T, Tibshirani R and others 2000 Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors) *Ann. Stat.* **28** 337–407
[7] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H 2009 The WEKA data mining software: an update *ACM SIGKDD Explor. Newsl.* **11** 10–8
[8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V and others 2011 Scikit-learn: Machine learning in Python *J. Mach. Learn. Res.* **12** 2825–2830
[9] Deelder W, Christakoudi S, Phelan J, Diez Benavente E, Campino S, McNerney R, Palla L and Clark T G 2019 Machine learning predicts accurately Mycobacterium tuberculosis drug resistance from whole genome sequencing data *Front. Genet.* **10** 922
[10] Yang Y, Niehaus K E, Walker T M, Iqbal Z, Walker A S, Wilson D J, Peto T E A, Crook D W, Smith E G, Zhu T and others 2018 Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data *Bioinformatics* **34** 1666–1671
[11] Kouchaki S, Yang Y, Walker T M, Sarah Walker A, Wilson D J, Peto T E A, Crook D W, Consortium Cr and Clifton D A 2019 Application of machine learning techniques to tuberculosis drug resistance analysis *Bioinformatics* **35** 2276–2282
[12] Chen M L, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane I S, Beam A and Farhat M 2018 Deep learning predicts tuberculosis drug resistance status from whole-genome sequencing data *BioRxiv* 275628
[13] Hicks A L, Wheeler N, Sánchez-Busó L, Rakeman J L, Harris S R and Grad Y H 2019 Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data *PLoS Comput. Biol.* **15** e1007349
[14] Jamal S, Khubaib M, Gangwar R, Grover S, Grover A and Hasnain S E 2020 Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis *Sci. Rep.* **10** 1–16
[15] Chen M L, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, Kohane I S, Beam A and Farhat M 2019 Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction *EBioMedicine* **43** 356–369
[16] Huang B F F and Boutros P C 2016 The parameter sensitivity of random forests *BMC Bioinformatics* **17** 331

Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data

ORIGINALITY REPORT

12%	8%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	media.neliti.com Internet Source	2%
2	academic-accelerator.com Internet Source	1%
3	Submitted to De La Salle University Student Paper	1%
4	Akanksha Sharma, Maria De Rosa, Neha Singla, Gurpal Singh, Ravi P. Barnwal, Ankur Pandey. "Tuberculosis: An Overview of the Immunogenic Response, Disease Progression, and Medicinal Chemistry Efforts in the Last Decade toward the Development of Potential Drugs for Extensively Drug-Resistant Tuberculosis Strains", Journal of Medicinal Chemistry, 2021 Publication	1%
5	Lecture Notes in Computer Science, 2013. Publication	1%
6	Submitted to University of Liverpool Student Paper	

		1 %
7	www.researchgate.net Internet Source	1 %
8	cg-korea.or.kr Internet Source	1 %
9	www.science.gov Internet Source	1 %
10	Priscila Lamb Wink, Zilpa Adriana Sanchez Quitian, Leonardo Astolfi Rosado, Valnes da Silva Rodrigues et al. "Biochemical characterization of recombinant nucleoside hydrolase from Mycobacterium tuberculosis H37Rv", Archives of Biochemistry and Biophysics, 2013 Publication	1 %
11	link.springer.com Internet Source	1 %
12	livrepository.liverpool.ac.uk Internet Source	1 %
13	Muchamad Taufiq Anwar, Wiwien Hadikurniawati, Edy Winarno, Wahyu Widiyatmoko. "Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia", 2020 3rd International Seminar on Research of Information	1 %

Technology and Intelligent Systems (ISRITI), 2020
Publication

14	aclweb.org Internet Source	1 %
15	Abhinav Sharma, Edson Machado, Karla Valeria Batista Lima, Philip Noel Suffys, Emilyn Costa Conceição. "Tuberculosis drug resistance profiling based on machine learning: A literature review", The Brazilian Journal of Infectious Diseases, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		