Indonesian Journal of Medical Chemistry and Bioinformatics

Volume 2 | Number 1 Article 4

10-9-2023

Metabolomic Insights into Tuberculosis: Machine Learning Approaches for Biomarker Identification

Miftahul Khair Akbar
Universitas Indonesia, Jakarta, miftahul.khair@ui.ac.id

Arief Aulia Rahman
Permata Cibubur Hospital, West Java, ariefauliarahman@gmail.com

Recommended Citation
Khair Akbar, Miftahul and Aulia Rahman, Arief (2023) "Metabolomic Insights into Tuberculosis: Machine Learning Approaches for Biomarker Identification," *Indonesian Journal of Medical Chemistry and Bioinformatics*: Vol. 2: No. 1, Article 4.
DOI: 10.7454/ijmcb.v2i1.1020
Available at: https://scholarhub.ui.ac.id/ijmcb/vol2/iss1/4

INDONESIAN JOURNAL OF MEDICAL CHEMISTRY AND BIOINFORMATICS

Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia

Article

Metabolomic Insights into Tuberculosis: Machine Learning Approaches for Biomarker Identification

Miftahul Khair Akbar [1]*, Arief Aulia Rahman [2]

[1] Magister Program in Biomedical Science, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia, 10430; miftahulk04@gmail.com
[2] General Practitioner in Permata Cibubur Hospital, Bekasi, West Java, Indonesia, 17435; ariefauliarahman@gmail.com
* Correspondence: miftahulk04@gmail.com

Abstract: The lung parenchyma is largely impacted by the infectious condition known as pulmonary tuberculosis (pulmonary TB) when the immune system creates a wall around the germs in the lungs, a tiny, hard bulge known as a tubercle develops, earning the disease the name tuberculosis. Although the majority of TB germs target the lungs, they can also harm other bodily organs. The identification of TB biomarkers, which are crucial for diagnosis, treatment monitoring, risk analysis, and prognosis, has been the subject of extensive research. Differences in metabolites between normal cells and tuberculosis are considered to be able to support the diagnosis of tuberculosis. Metabolite data was taken from the Metabolomic workbench and further identification and prediction were carried out in silico. A total of 44 samples found 69 metabolites which were then carried out further analysis. Found as many as 5 metabolites that play an important role in tuberculosis. Of the 5 metabolites, 2 candidate biomarkers were found which are known to have potential as biomarkers. The candidate biomarkers for these metabolites are trans-3-methyluric acid and nicotinic acid. However, this simulation needs further testing to obtain more accurate biomarkers and support the diagnosis.

Keywords: Biomarker; Tuberculosis; Metabolite

Citation: Akbar, M.K; Rahman, A.A. Metabolomic Insights into Tuberculosis: Machine Learning Approaches for Biomarker Identification. *Ind. J. Med. Chem. Bio. IJMCB.* **2023**, *2*, 1.

Received: Wed Jun 07, 2023
Accepted: Tue Oct 03, 2023
Published: Mon Oct 09, 2023

Publisher's Note: IJMCB stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Even though it is treatable, the disease tuberculosis (TB), which is brought on by the bacteria *Mycobacterium tuberculosis* (Mtb), is still rampant throughout the world. WHO estimates that in 2020, 10 million new cases of TB were reported, and 1.5 million people died from it, making it the second most lethal infectious disease after COVID-19 [1,2]. An infectious condition known as tuberculosis (TB) mostly impacts the lung parenchyma. The word tubercle, which describes small, hard nodules that develop as the immune system constructs a wall around the bacteria in the lungs, is the source of the term tuberculosis. TB is a chronic illness that frequently results in the growth of granulomas and tissue necrosis. When someone with active TB coughs, sneezes, or speaks, TB can spread through the air. The TB bacteria, *Mycobacterium tuberculosis*, is responsible for the direct communicable disease known as tuberculosis. Although the bulk of TB germs assault the lungs, they can also harm other body organs [3,4].

People who have positive acid-fast bacilli (AFB) tuberculosis during coughing or sneezing are the source of tuberculosis transmission. Patients sputum droplets, which contain microorganisms, are released into the air. Droplets containing the bacteria can linger in the air for several hours at room temperature [5]. If someone breathes in these droplets and gets an infection in their respiratory system. The bacteria that cause

DOI: https://doi.org/10.7454/ijmcb.v2i1.1020. E-ISSN : 2963-3818

tuberculosis can travel from the lungs to other body parts via the circulation, respiratory tract, or direct dissemination to other body parts after entering the human body through respiration. The quantity of bacteria a patient's lungs release determines how contagious they are [6]. The more positive the sputum test results, the more contagious the patient is deemed to be. The patient is regarded as non-infectious if the results of the sputum examination are negative (no visible germs). The quantity of droplets in the air and the length of time an individual is exposed to those droplets decide whether they infect them with tuberculosis [7].

Governments and scientific communities worldwide have made attempts to combat TB due to its enormous effects on public health. By strongly supporting basic and clinical research on tuberculosis, all United Nations members have pledged to eradicate the TB epidemic by 2030. The identification of TB biomarkers, which are crucial for diagnosis, treatment monitoring, risk analysis, and prognosis, has been the subject of extensive research [8]. Machine learning has been extensively used in biomarker development in recent years. When training a model to learn from data for a specific task, machine learning employs mathematical techniques [9]. Classification and feature selection are essential machine learning approaches for the discovery of biomarkers. A type of supervised learning called classification involves feeding the algorithm with labeled examples, each of which is represented by a set of features. Learning a function that can infer a sample's label from its features is the task at hand. [10]. In our case, machine learning is used to predict biomarkers using metabolites.

2. Results

PCA is a technique for extracting patterns from large amounts of complex data. The entered dataset will be transformed into two datasets: one with individual weights and one with principal component weights [12]. Figure 1 depicts graph analysis components in the form.

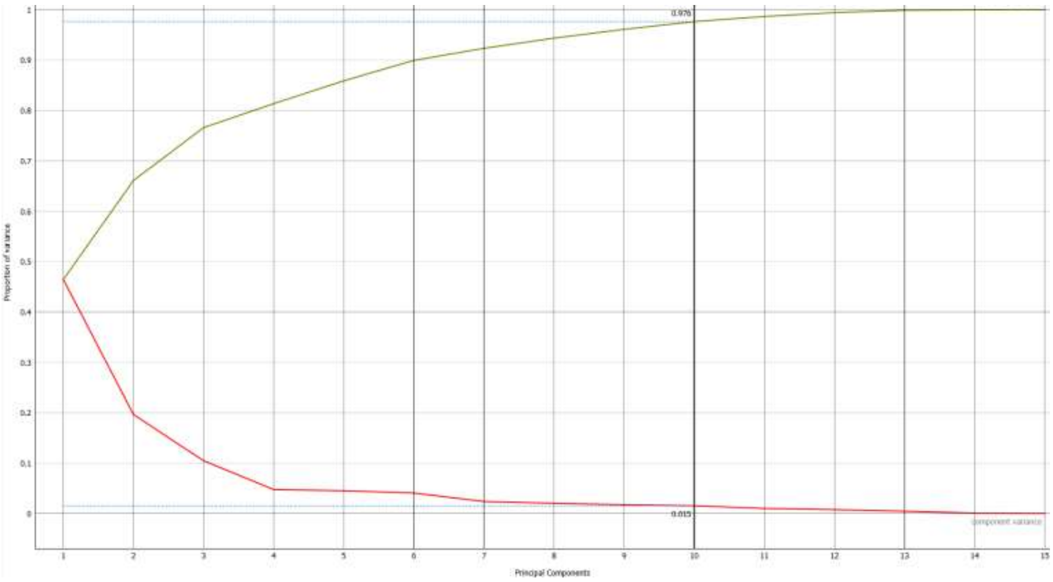


Figure 1. Scree diagram of tuberculosis with controls

A scree diagram is a diagram that is used to determine the success of PCA in a dataset. The main components will be created based on the number of possible variations. PC1 will cover more variants than PC2 in the hierarchy. A scree diagram can assist in determining how much variation each PC (principal component) can capture. The y-axis represents the eigenvalues, while the x-axis represents the principal components. The eigenvalues will show how much variation there is. The ideal curve is steep and then sharply bends; the bend is the point of intersection, and the curve flattens. In Figure 1, the red line represents the variance covered per component, while the green line represents the total variance covered by the components. The red line in figure 1 is a curved scree. The curve will drop sharply at first, then bend since the first group is usually highly variable. Group 2 will have less variability because it is not excessive, whereas group 3 and subsequent variations will have less variability, resulting in a flatter curve [13]. The metabolomic tuberculosis dataset with controls can be distinguished from the scree diagram in Figure 1. The obvious bend in the principal component 2 reflected by the red line demonstrates this. Figure 2 shows a scatter plot that depicts the difference between metabolomic tuberculosis and controls.

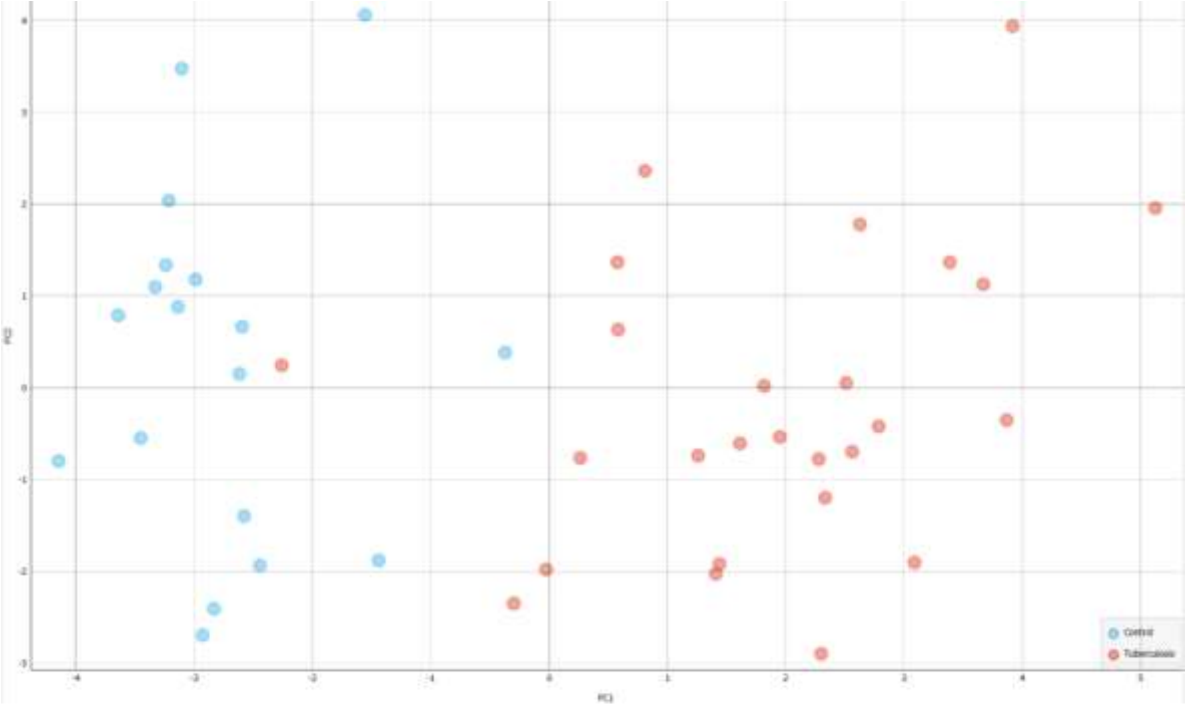
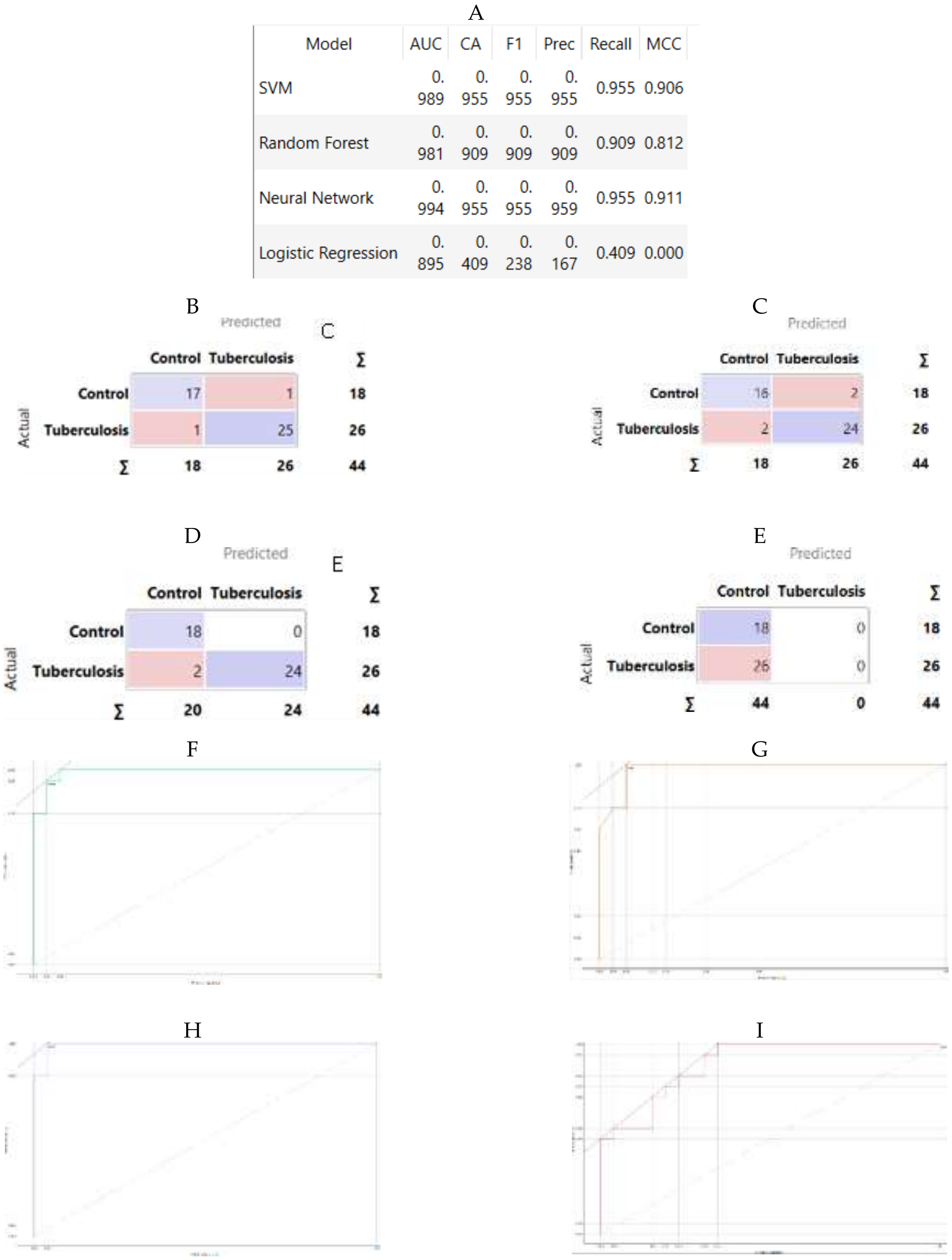


Figure 2. Scatter Plot of Tuberculosis Metabolomics Dataset with Controls

In Figure 2, there is a clear distinction between PC1 and PC2. The graph also shows that dot separation occurs among patients with tuberculosis and control. It can be demonstrated that there is a difference in metabolite from tuberculosis patients and control. Prediction of these differences can be accomplished through the use of machine learning. Commonly used machine learning techniques include SVM, random forest, neural networks, and logistic regression. The results of machine learning prediction can be seen in Figure 3.

Figure 3 shows that the accuracy, specificity, and sensitivity from every method. All methods have an AUC value > 70%. AUC of 70% indicates that the data can be accurately, specifically, and sensitively analyzed and compared. Neural network method has the best AUC value with an AUC value of 0,994. The ROC graph supports this, with y axis representing sensitivity and x axis representing specificity. If the sensitivity rises while the specificity falls, the outcome will improve [14].

DOI: https://doi.org/10.7454/ijmcb.v2i1.1020. E-ISSN : 2963-3818



Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0. 989	0. 955	0. 955	0. 955	0.955	0.906
Random Forest	0. 981	0. 909	0. 909	0. 909	0.909	0.812
Neural Network	0. 994	0. 955	0. 955	0. 959	0.955	0.911
Logistic Regression	0. 895	0. 409	0. 238	0. 167	0.409	0.000

	Control	Tuberculosis	Σ
Control	17	1	18
Tuberculosis	1	25	26
Σ	18	26	44

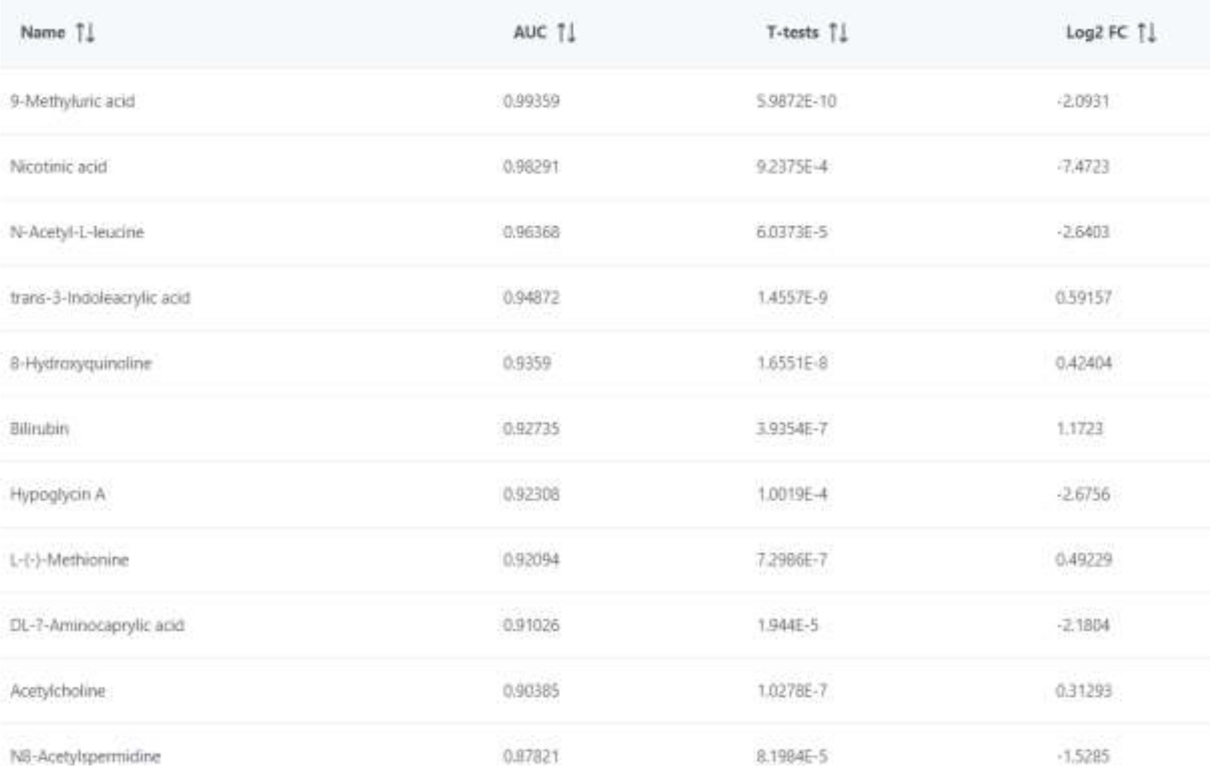
	Control	Tuberculosis	Σ
Control	16	2	18
Tuberculosis	2	24	26
Σ	18	26	44

	Control	Tuberculosis	Σ
Control	18	0	18
Tuberculosis	2	24	26
Σ	20	24	44

	Control	Tuberculosis	Σ
Control	18	0	18
Tuberculosis	26	0	26
Σ	44	0	44

Figure 3. Prediction of Machine Learning (A) Prediction of Machine Learning (B) Confussion matrix SVM (C) Confussion matrix Random Forest (D) Confussion matrix Neural Network (E) Confussion matrix Logistic Regression (F) ROC curve SVM (G) ROC Curve Random Forest (H) ROC Curve Neural Network (I) ROC Curve Logistic Regression

According to the results of the statistical data analysis, metabolite in patients with tuberculosis can be compared to metabolite control. Of the 69 metabolites data, 5 metabolites have the potential to be used as biomarkers for tuberculosis. The metabolite data obtained using Orange Software was compared using MetaboAnalyst 5.0. The data obtained using MetaboAnalyst 5.0 is the same as the data obtained from Orange Software. Metabolites can be found in Figure 4.



Name ↑↓	AUC ↑↓	T-tests ↑↓	Log2 FC ↑↓
9-Methyluric acid	0.99359	5.9872E-10	-2.0931
Nicotinic acid	0.98291	9.2375E-4	-7.4723
N-Acetyl-L-leucine	0.96368	6.0373E-5	-2.6403
trans-3-Indoleacrylic acid	0.94872	1.4557E-9	0.59157
8-Hydroxyquinoline	0.9359	1.6551E-8	0.42404
Bilirubin	0.92735	3.9354E-7	1.1723
Hypoglycin A	0.92308	1.0019E-4	-2.6756
L-(-)-Methionine	0.92094	7.2986E-7	0.49229
DL-?-Aminocaprylic acid	0.91026	1.944E-5	-2.1804
Acetylcholine	0.90385	1.0278E-7	0.31293
N8-Acetylspermidine	0.87821	8.1984E-5	-1.5285

Figure 4. Metabolites obtained from MetaboAnalyst 5.0

Figure 4 is the result of the analysis using MetaboAnalyst 5.0. The results obtained show a list of metabolites that can be used as biomarkers in tuberculosis. Based on research by Ray et al (2010), the area under curve (AUC) value of 0.75 – 0.90 is classified as good diagnostic value [15]. There are 5 metabolites obtained from Orange Software and MetaboAnalyst 5.0 which can be used as biomarkers for tuberculosis namely 9-metyluric acid, nicotinic acid, N-acetyl-L-leucine, trans-3-indoleacrylic acid, and 8-Hydroxyquinoline. These 5 metabolites have an AUC value of > 0.9.

3. Discussion

Machine learning is used to determine the accuracy and specificity of data that has been collected [16]. SVM (support vector machine) is an algorithm used to compress data in N-dimensional space using hyperplanes [17]. Supervised learning is a subset of machine learning that includes logistic regression. In order to determine the likelihood of two variables in a specific situation, logistic regression is widely utilized. [18]. A classification technique called random forest looks at a dataset's several interesting points. In contrast to opinion polls, the random forest algorithm was used to generate opinion polls. This will have a negative impact on the outcome of the prediction. Random forest has a distinct advantage in that it performs simulations more quickly than other algorithms and can make decisions with high accuracy [19]. A neural network is an artificial intelligence (AI) method that teaches computers to think like humans. The human brain, which synergizes to create memories, inspired neural network learning. Neural networks can help solve complex problems with high sensitivity [20]. A confusion matrix is a table that is used to determine a piece of data's accuracy, specificity, and sensitivity [21].

Trans-3-indoleacrylic acid is a tryptophan metabolite which is related to tuberculosis. Trans-3- indoleacrylic acid as a biomarker for TB lies in its association with tryptophan metabolism, which can be altered during TB infection. Tryptophan is an essential amino acid that is metabolized through various pathways, including the kynurenine pathway. During TB infection, there may be changes in tryptophan metabolism, resulting in the production of metabolites like trans-3-indoleacrylic acid. The proposed mechanism is that *Mycobacterium tuberculosis*, the bacteria causing TB, can induce the production of indoleamine 2,3-dioxygenase (IDO), an enzyme involved in tryptophan metabolism. Increased IDO activity can lead to alterations in the kynurenine pathway, resulting in the accumulation of trans-3- indoleacrylic acid [22].

We are unaware of any connections between TB and the other three distinct metabolites: 9- methyluric acid, N-acetyl-L-leucine, and 8-hydroxyquinoline. A methyl derivative of uric acid called 9- methyluric acid is infrequently detected in human urine. One of the purine elements in urinary calculi is 9-methyluracil. Methylated purines are produced when methylxanthines like caffeine, theophylline, and theobromine are metabolized. Its decreased blood levels were connected to the metabolic syndrome and prediabetes [23]. An acetylated derivative of leucine known as N-acetyl-L-leucine is produced by a particular N-acetyltransferase or as a result of the proteolytic breakdown of N-acetylated proteins. It is employed to treat neurological disorders [24]. 8-Quinolinol is a straightforward alkaloid with a wide range of biological effects. 8-quinolinol is thought to be one of the most effective treatments for various neurological illnesses that are brought on by too much ROS [25]. 8-hydroxyquinolines have been found to be bactericidal against *Mycobacterium tuberculosis* [26].

Nicotinic acid is produced in large enough quantities to serve as the foundation of the niacin test used to distinguish *M. tuberculosis* from other mycobacterial species, nicotinic acid is known to play a significant part in oxidation reduction processes in mycobacterial metabolism. It is generally known that

M. tuberculosis generates nicotinic acid in a lab setting. It was discovered that tuberculous patient total nicotinic acid readings were noticeably greater than those of normal persons [27].

4. Materials and Methods

4.1 Search for Metabolite Data in Metabolomic Workbench

The Metabolomic Workbench can be found at https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR001562. The keyword tuberculosis will be entered in the query section. Following that, the appropriate data search is performed, where it can be seen in the "view" section whether there is metabolite data or only meta data. To reduce bias, it is also necessary to sort the species that are chosen from humans rather than test animals in the metadata section. The time of data collection is also an important parameter to consider because the more recent the year of publication, the more recent the technology used, resulting in a renewable metabolite profile.

4.2 Biomarker Data Analysis using Orange Data Mining

The data obtained from the metabolomic workbench is in .txt. The data must be converted to.xls format before it can be transferred and analyzed in orange data mining. Sample data should be organized in columns, while metabolic data should be organized in rows. The data is then imported into the Orange software and normalized before the analysis begins. The normalized data is then ranked in order, followed by principal component analysis (PCA). PCA visualization is accomplished through the use of scatter plots.

PCA has several benefits in machine learning including lack of redundancy of data given the orthogonal components, reduced complexity in images' grouping with the use of PCA, smaller database representation since only the trainee images are stored in the form of their projections on a reduced basis, and reduction of noise since the maximum variation basis is chosen and so the small variations in the background are ignored automatically [11].

4.3 Machine Learning Prediction using Orange Data Mining

Machine learning is used to predict orange data that has been analyzed. Support vector machines (SVM), logistic regression, and neural algorithms are examples of machine learning. The AUC curve and the confusion matrix are visible parameters. Data prediction will be obtained from the machine learning algorithm by examining the sensitivity and specificity.

4.4 Metabolite analysis using MetaboAnalyst 5.0

Metabolite data analyze using Metaboanalyst 5.0 (https://www.metaboanalyst.ca/MetaboAnalyst/). The previously obtained data is inputted into the 'biomarker analysis' section. Then normalization by median and ROC analysis were carried out. The list of metabolites obtained is then interpreted.

This paper uses metabolites data from Metabolomic workbench using Project ID PR001652 and Study ID ST002428. Metabolite data was obtained from human serum and was characterized using MS method. The search for biomarkers of these metabolites is intended to result in a diagnostic kit that will be used by humans in the future. As a result, the species *Homo sapiens* is used to avoid data bias.

New information on the metabolic processes and cellular physiology of living things is being gained by the use of mass spectrometry (MS) to metabolite analysis. These methods are increasingly being employed to find biomarkers and study metabolic dynamics systematically. Untargeted metabolomic profiling enables the detection of unexpected compounds, whereas targeted techniques are used to monitor and frequently quantify certain metabolites of interest.

Metabolomic workbench data were then prepared for opening in the orange analysis window. The sample must be in the column section, while the metabolites must be in the row section, according to the data criteria. Figure 5 depicts a data processing step performed with orange software.

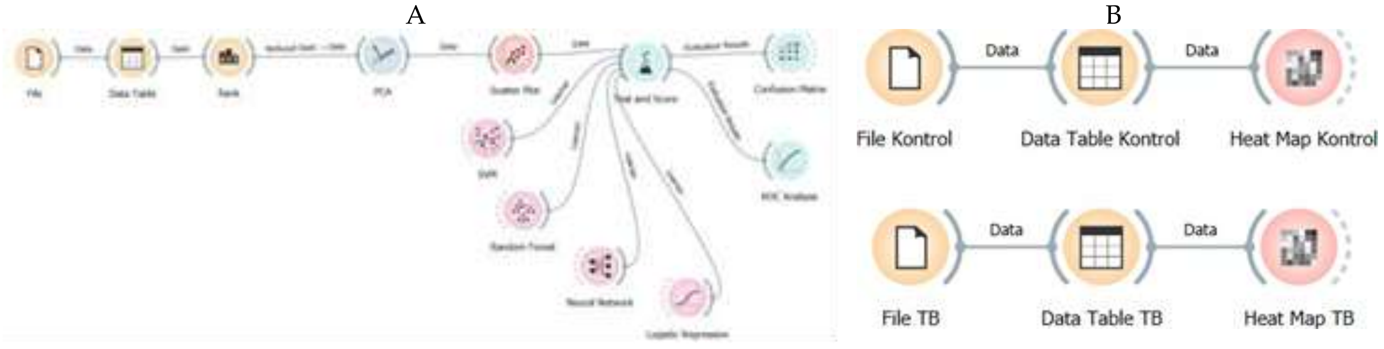


Figure 5. Data processing steps in orange (A) Machine learning processing and prediction steps (B) Heatmap data processing steps.

5. Conclusions

5 metabolites were identified as the most significant in differentiating the sample group marked with an AUC value > 0.9, namely 9-methyluric acid, nicotinic acid, trans-3-

indoleacrylic acid, N-acetyl- L-leucine, and 8-Hydroxyquinoline. Of the five metabolites, some metabolites that has been studied is related to tuberculosis, namely trans-3-indoleacrylic acid. Trans-3-indoleacrylic acid is related to indoleamine 2,3-dioxygenase activity in tuberculosis. Nicotinic acid is produced in sufficient quantities to serve as the foundation of the niacin test used to distinguish *M. tuberculosis*, and it is known to play a significant role in oxidation-reduction events in mycobacterial metabolism. The results of the analysis can be used as material for consideration in the selection of tuberculosis biomarkers, but further studies are needed to obtain more accurate biomarkers and to further examine the association of the metabolites obtained from the analysis results.

References

1. WHO. Global Tuberculosis Report 2019; WHO: Geneva, Switzerland, 2019.
2. WHO. Global Tuberculosis Report 2021; WHO: Geneva, Switzerland, 2021.
3. CDC. Guidelines for prevention and treatment of opportunistic infections in HIV-infected adults and adolescents. MMWR 2009; 58 (No. RR-4). www.cdc.gov/mmwr/preview/mmwrhtml/rr58e324a1.htm?s_cid=rr58e324a1_e
4. American Thoracic Society and CDC. Diagnostic standards and classification of tuberculosis in adults and children. Am J Respir Crit Care Med 2000; 161 (4): 1376– 1395. http://ajrccm.atsjournals.org/cgi/reprint/161/4/1376
5. CDC. A strategic plan for the elimination of tuberculosis from the United States. MMWR 1989; 38 (Suppl No. S-3). www.cdc.gov/mmwr/preview/mmwrhtml/00001375.htm
6. American Thoracic Society and Infectious Diseases Society of America. Diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. Am J Respir Crit Care Med 2007; 175: 367–416. www.thoracic.org/statements/resources/mtpi/nontuberculous-mycobacterial-diseases.pdf
7. CDC. Essential components of a tuberculosis prevention and control program: Recommendations of the Advisory Council for the Elimination of Tuberculosis. MMWR 1995; 44 (No. RR-11). www.cdc.gov/mmwr/preview/mmwrhtml/00038823.htm
8. Yu, Y., Jiang, X. X., & Li, J. C. (2023). Biomarker discovery for tuberculosis using metabolomics. *Frontiers in molecular biosciences*, *10*, 1099654. https://doi.org/10.3389/fmolb.2023.1099654
9. Amalia, F., Syamsunarno, M. R. A. A., Triatin, R. D., Fatimah, S. N., Chaidir, L., & Achmad, T. H. (2022). The Role of Amino Acids in Tuberculosis Infection: A Literature Review. *Metabolites*, *12*(10), 933. https://doi.org/10.3390/metabo12100933
10. I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202
11. R. D. Ledesma, P. Valero-Mora, and G. Macbeth, "The Scree Test and the Number of Factors: a Dynamic Graphics Approach," Span. J. Psychol., vol. 18, no. June, p. E11, 2015, doi: 10.1017/sjp.2015.13
12. P. Larrañaga et al., "Machine learning in bioinformatics," Brief. Bioinform., vol. 7, no. 1, pp. 86–112, 2006, doi: 10.1093/bib/bbk007.
13. H. Ahmadi, P. Pichappan, and E. Ariwa, Communications in Computer and Information Science: Preface, vol. 241 CCIS. 2011
14. C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," J. Educ. Res., vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.
15. A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," Indones. J. Artif. Intell. Data Min., vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.
16. F. Dhimas Syahfitra, R. Syahputra, and K. Trinanda Putra, "Implementation of Backpropagation Artificial Neural Network as a Forecasting System of Power Transformer Peak Load at Bumiayu Substation," J. Electr. Technol. UMY, vol. 1, no. 3, pp. 118–125, 2017, doi: 10.18196/jet.1316.
17. D. Visa Sofia, "Confusion Matrix-based Feature Selection Sofia Visa," ConfusionMatrix-based Featur. Sel. Sofia, vol. 710, no. January, p. 8, 2011.
18. K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Casp. J. Intern. Med., vol. 4, no. 2, pp. 627–635, 2013.
19. Ray P, Le Manach Y, Riou B, Houle TT. Statistical evaluation of a biomarker. Anesthesiology. 2010 Apr;112(4):1023-40. doi: 10.1097/ALN.0b013e3181d47604. PMID: 20234303.
20. Fernandes AF, Gonçalves LG, Bento M, Anjo SI, Manadas B, Barroso C, Villar M, Macedo R, Simões MJ, Coelho AV. Mass Spectrometry-Based Proteomic and Metabolomic Profiling of Serum Samples for Discovery and Validation of Tuberculosis Diagnostic Biomarker Signature. *International Journal of Molecular Sciences*. 2022; 23(22):13733. https://doi.org/10.3390/ijms232213733
21. Alsoud, L.O.; Soares, N.C.; Al-Hroub, H.M.; Mousa, M.; Kasabri, V.; Bulatova, N.; Suyagh, M.; Alzoubi, K.H.; El-Huneidi, W.; Abu-Irmaileh, B.; et al. Identification of Insulin Resistance Biomarkers in Metabolic Syndrome Detected by UHPLC-ESI-QTOF-MS. *Metabolites* 2022, *12*, 508.
22. Li, X.-L.; Li, W.; Qi, C.; Zhao, L.-L.; Jin, W.; Min, J.Z. Determination of N-acetyl-DL-leucine in the saliva of healthy volunteers and diabetic patients using ultra-performance liquid chromatography with fluorescence detection. *Clin. Chim. Acta* 2022, *526*, 66–73. [Google Scholar] [CrossRef]

23. Blake, O. A., Bennink, M. R., & Jackson, J. C. (2006). Ackee (Blighia sapida) hypoglycin A toxicity: dose response assessment in laboratory rats. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*, *44*(2), 207–213. https://doi.org/10.1016/j.fct.2005.07.002
24. Prachayasittikul, V., Prachayasittikul, S., Ruchirawat, S., & Prachayasittikul, V. (2013). 8-Hydroxyquinolines: a review of their metal chelating properties and medicinal applications. *Drug design, development and therapy*, *7*, 1157–1178. https://doi.org/10.2147/DDDT.S49763
25. Syhre, M., Manning, L., Phuanukoonnon, S., Harino, P., & Chambers, S. T. (2009). The scent of Mycobacterium tuberculosis--part II breath. *Tuberculosis (Edinburgh, Scotland)*, *89*(4), 263–266. https://doi.org/10.1016/j.tube.2009.04.003