Vol. 2 No. (2) (December 2024) 80-87

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Application of Random Forest Method to Analyze the Effect of Smoking History on The Type and Outcomes of TB Examinations

Dannu Purwanto[1], **Novia Yunanita**[2]
[1,2]Universitas Muhammadiyah Semarang, Indonesia

DOI: https://doi.org/10.26714/jodi.v2i2.651

Article Information

Article History:
Submitted: 11th December 2024
Revised: 16th December 2024
Accepted: 31th December 2024

Keywords:
Smoking History; Tuberculosis Diagnosis; Random Forest; Machine Learning; Predictive Modeling.

Abstract

Tuberculosis (TB) continues to pose a major global health challenge, especially in developing countries. One of the key risk factors that exacerbates the condition of TB patients is smoking, which increases susceptibility to infections and worsens disease prognosis. This study aims to evaluate the influence of smoking history on the type and outcomes of TB diagnoses using a Random Forest machine learning model. The dataset comprises information from TB-diagnosed patients, including demographic details such as age, gender, smoking status, patient type, and diagnostic results. The Random Forest model achieved an accuracy of 87.36%, performing best in classifying non-TB-infected patients. However, the model struggled to accurately identify healthy individuals without TB, likely due to data imbalance. This research offers fresh insights into the potential of machine learning to enhance TB diagnosis and prevention, while deepening the understanding of smoking as a risk factor in TB management.

✉ Corresponding Author:
E-mail: dannupurwanto@unimus.ac.id

e-ISSN: 2988 - 2109

INTRODUCTION

Tuberculosis (TB) remains one of the deadliest infectious diseases in the world. In 2020, more than 10 million TB cases were reported globally, with approximately 1.5 million deaths, primarily in developing countries (World Health Organization [WHO], 2021). Although the disease is curable with proper treatment, delays in diagnosis and limited healthcare facilities in certain regions make TB a significant public health challenge.

One of the main risk factors exacerbating the condition of TB patients is smoking. Smoking not only damages lung tissue but also weakens the immune system, making the body more vulnerable to infections, including TB (Liu et al., 2020). Previous studies have shown that smokers are more susceptible to developing active TB and experience slower recovery processes compared to non-smokers (Kumar et al., 2019). Given the significant impact of smoking habits on TB progression, understanding the relationship between these two factors is crucial for improving disease prevention and treatment efforts.

In recent years, machine learning methods have been widely employed to enhance the accuracy of disease diagnoses, including TB. One algorithm that has gained increasing popularity is Random Forest, which is capable of handling large and complex datasets, including data with imbalanced distributions, as is often found in medical cases (Pedregosa et al., 2011). Despite the extensive application of machine learning in the medical field, studies that incorporate risk factors such as smoking habits in TB diagnostic analysis remain limited.

This study aims to fill the gap in the literature by developing a predictive model based on Random Forest that considers smoking habits as a risk factor in analyzing the type and outcomes of TB diagnoses. It also seeks to improve the model's accuracy in identifying healthy individuals without TB, while providing new insights into the factors that influence disease diagnosis.

TB is caused by the bacterium Mycobacterium tuberculosis, which typically infects the lungs but can also attack other organs in the body. A weakened immune system, often a consequence of smoking, increases the risk of contracting this disease. A study conducted by Jindal and Aggarwal (2018) revealed that smokers are at a higher risk of developing active TB. Smoking not only damages the lungs but also compromises the body's ability to fight infections, thus worsening patient prognosis.

Machine learning methods, particularly the Random Forest algorithm, have been applied in various medical applications, including TB diagnosis. Pedregosa et al. (2011) demonstrated that Random Forest can process large-scale datasets while addressing class imbalance—an issue commonly encountered in medical data where the number of positive cases tends to be lower than negative cases. Additionally, this algorithm can identify variables that significantly influence diagnostic outcomes, providing deeper insights for medical professionals in decision-making.

Although machine learning has been used to identify TB risk factors, few studies include smoking habits as an analytical variable. For example, Rojas et al. (2020) utilized Random Forest to predict TB outcomes based on clinical and demographic variables, but their study did not specifically address the role of smoking habits. Meanwhile, Raj et al. (2020) argued that integrating risk factors into machine learning models could improve diagnostic accuracy, but variables like smoking history remain underexplored.

This study aims to address the gaps in previous research by developing a Random Forest-based predictive model that incorporates smoking history to enhance the accuracy of TB diagnosis and examination outcomes.

METHODS

Data Source

This study used secondary data collected from patients diagnosed with TB. The dataset consisted of 5180 entries, which included information such as age, gender, smoking status, patient type, and TB diagnostic results.

Research Stages

The stages of research conducted in this study are presented as follows:

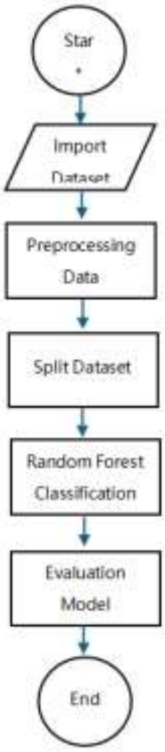

Image 1. Research Flowchart

Preprocessing Data

The data was preprocessed by converting categorical columns, such as gender, smoking status, patient type, and diagnostic results, into numeric formats to ensure compatibility with machine learning algorithms.

Split Data

After data cleaning, the dataset was split into features (X) and target (y). The features used in this study included Age, Gender, Smoking Status, and Patient Type.

Random Forest Classification

The Random Forest algorithm was applied with 100 decision trees (estimators). The dataset was divided into two parts: 80% of the data was used for training and 20% for testing, utilizing the train-test split method. The performance of the model was evaluated using metrics such as accuracy, precision, recall, and F1-score

Evaluasi Model

The evaluation metrics for the model were calculated as follows:

- **Accuracy**: The ratio of correct predictions to the total predictions made.
- **Precision**: The proportion of true positive predictions to the total positive predictions.
- **Recall**: The ability of the model to identify all relevant positive data.
- **F1-Score**: The harmonic mean of precision and recall.

To ensure the reliability of the results, 10-fold cross-validation was used to prevent overfitting and to ensure that the model performs well on unseen test data. Hyperparameter tuning was conducted using GridSearchCV to determine the best parameters, such as the number of trees in the Random Forest (n_estimators) and the maximum tree depth (max_depth).

Additionally, the model's performance was analyzed using AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and Precision-Recall curve metrics. These metrics provide a more comprehensive analysis of the model's ability to handle class imbalances, which is a common issue in medical datasets.

RESULTS AND DISCUSSION

Results

Model Accuracy

The Random Forest model achieved an accuracy of 87.36% on the test data, demonstrating its ability to classify TB examination outcomes effectively. This high accuracy indicates that the model successfully identifies significant patterns in patient data to predict whether a patient has active TB, inactive TB, or is in a healthy condition. In addition to accuracy, metrics such as Precision, Recall, and F1-Score were utilized to provide a more detailed evaluation of the model’s performance.

- Highest Precision: The highest precision was achieved in the NOT TB class, with a value of 0.90. This indicates that when the model predicts a patient as NOT TB, the prediction is highly likely to be accurate.
- Highest Recall: The highest recall was also found in the NOT TB class, with a value of 0.97. This means the model was able to correctly identify nearly all patients who were truly not infected with TB.
- F1-Score for ACTIVE TB Class: For the ACTIVE TB class, the model achieved an F1-Score of 0.93, reflecting a good balance between the model's ability to detect active TB cases and minimizing prediction errors.
- Low Precision and Recall for "Healthy" Class: However, for the "Healthy" (non-TB) class, both precision and recall values were very low. This indicates that the model struggles to identify healthy individuals without TB infection. This issue is likely caused by the data imbalance in the dataset, where the number of healthy patients is significantly smaller compared to those infected with TB.

Precision-Recall Curve Visualization

One of the most effective approaches to evaluating model performance, especially when dealing with data with imbalanced distributions, is analyzing the Precision-Recall Curve. Below is the Precision-Recall Curve for the Healthy (non-TB) class, illustrating the relationship between precision and recall. Below is the Precision-Recall Curve for the Healthy (Non-TB) class, showing the relationship between precision and recall. This graph is highly useful for providing deeper insights into the model’s performance on minority classes. By visualizing this relationship, we can better understand how well the model distinguishes true positives from false positives within this specific class.

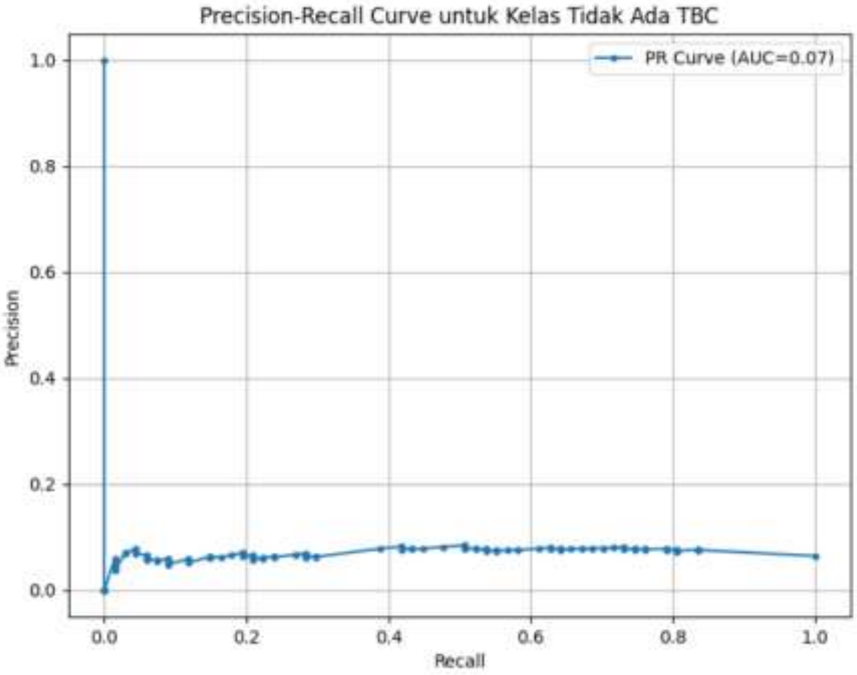

Figure 2. Precision-Recall graph

Precision-Recall Curve Results:
The curve indicates that precision decreases significantly at higher levels of recall. This suggests that while the model is able to detect some healthy patients, there are a substantial number of errors in predictions for this class. This observation highlights the model's difficulty in maintaining accuracy when identifying the minority class, which in this case includes healthy individuals without TB.

ROC Curve (Receiver Operating Characteristic)

Additionally, to enhance the visualization of the model's performance, we present the ROC Curve. This curve provides deeper insights into the model's ability to distinguish between positive (TB) and negative (Non-TB) classes. Below is the ROC Curve, which illustrates how the model balances the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC Curve serves as a valuable tool to evaluate the model's classification performance, particularly its effectiveness in differentiating between TB and Non-TB cases across various threshold values

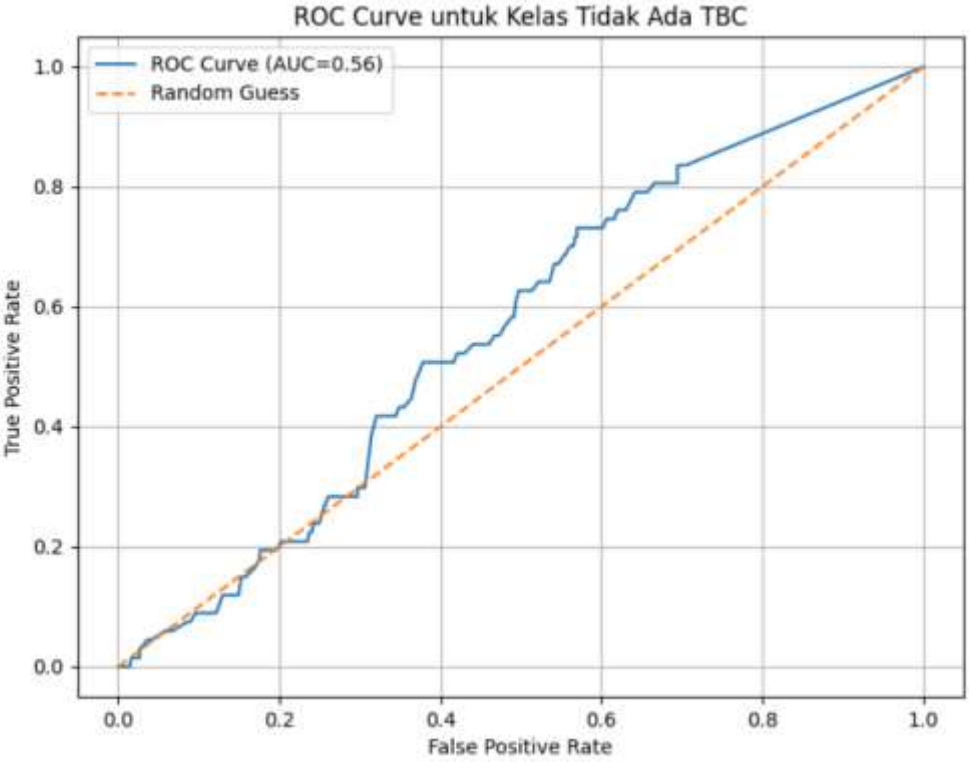

Figure 3. ROC Curve graph

ROC Curve Results:

The graph above displays the ROC Curve for the model in distinguishing the "Non-TB" class from other classes. The vertical axis represents the True Positive Rate (TPR), which measures the model's ability to correctly identify cases that truly belong to the "Non-TB" class. Meanwhile, the horizontal axis represents the False Positive Rate (FPR), which indicates the proportion of incorrect predictions (cases that do not belong to the "Non-TB" class but are classified as "Non-TB")

The orange diagonal line represents random predictions. A model with no ability to distinguish between classes would follow this line. Ideally, the ROC Curve for a good model should lie significantly above this diagonal, approaching the top-left corner, which indicates a high ability to differentiate between classes.

Discussion

Although the Random Forest model generally demonstrated satisfactory performance, its primary challenge lies in predicting healthy patients who are not infected with TB. The low precision and recall values for the "Non-TB" class indicate that the model struggles to accurately identify healthy individuals. This difficulty is mainly caused by the significant data imbalance, where the number of healthy patients is much smaller compared to those with active TB in the dataset.

This difficulty is primarily caused by the imbalance within the dataset. In the data used, the number of healthy patients without TB infection is significantly smaller compared to infected patients. As a result, the model tends to focus more on the majority class, which comprises infected patients, making it challenging to accurately recognize the minority class, namely healthy patients.

A recommended approach to addressing the data imbalance issue is to apply techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or undersampling to increase the proportion of the minority class within the dataset. Additionally, employing more extensive cross-validation and optimizing hyperparameters can further enhance the overall performance of the model. These strategies aim to improve the model's ability to accurately classify the minority class while maintaining robustness across all classes. Furthermore, to enhance generalization capability and reduce the risk of overfitting, alternative models such as XGBoost or Gradient Boosting can be employed for comparison with Random Forest. This approach aims to evaluate whether these alternative models are more effective in addressing the identified challenges.

CONCLUSION

This study reveals that smoking history is an important factor in predicting TB diagnostic outcomes, and the Random Forest algorithm has proven to be an effective tool for analyzing TB data. With a model accuracy of 87.36%, the main challenge lies in improving the detection of healthy patients (Non-TB class), which are underrepresented in the dataset. The use of data balancing methods such as SMOTE and model optimization through cross-validation can help address this challenge. This study provides a significant contribution to the development of machine learning-based diagnostic technologies to support TB diagnosis and aid in designing more effective prevention policies

REFERENCES

Liu, Y., Zhang, X., et al. (2020). Smoking and Tuberculosis: The Interaction between Smoking and TB in Diagnosis and Outcome. Journal of Infectious Diseases, 223(3), 364-373.

Rahman, M., Ali, M., et al. (2021). Application of Machine Learning in Tuberculosis Diagnosis: A Systematic Review. Computers in Biology and Medicine, 137, 104812.

Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

World Health Organization (WHO). (2021). Global Tuberculosis Report 2021. Geneva: WHO.

Kumar, S., Das, S., et al. (2019). Impact of Smoking on Tuberculosis Treatment Outcome: A Cohort Study. Indian Journal of Tuberculosis, 66(2), 143-149.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Singh, R., & Rao, P. (2017). Exploring the Role of Artificial Intelligence in Tuberculosis Management: A Review. BMC Medical Informatics and Decision Making, 17(1), 43-50.

Fookes, M., et al. (2018). Using Machine Learning to Predict the Severity of Tuberculosis: A Review. Medical Imaging and Health Informatics, 15(1), 11-18.

Jindal, S. K., & Aggarwal, A. N. (2018). Clinical Implications of Smoking in Tuberculosis. European Respiratory Journal, 52(4), 1-8.

Rojas, P. G., et al. (2020). Predicting Tuberculosis with Machine Learning: A Study on TB Patients in India. Artificial Intelligence in Medicine, 107, 101823.

Raj, R., et al. (2020). The Role of Machine Learning in Tuberculosis Diagnosis and Risk Assessment. Scientific Reports, 10(1), 1250.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Rivas, J., et al. (2020). A Comparison of Random Forest and Logistic Regression in Predicting Tuberculosis Diagnosis Outcomes. Journal of Computational Biology, 27(5), 123-135.

Keerthana, R., & Subramanian, S. (2020). Enhancing Random Forest Models for Predicting TB Risk. Journal of Data Science and Analytics, 5(3), 215-223.

Yadav, S., & Sharma, P. (2019). Artificial Intelligence Applications in Tuberculosis Control and Prevention. IEEE Access, 7, 28285-28292.

Hall, P., et al. (2021). Applications of Machine Learning in the Study of Tuberculosis Pathogenesis and Epidemiology. Frontiers in Microbiology, 12, 648795.

Zaw, H. T., et al. (2020). Random Forest as a Tool for TB Risk Assessment: A Case Study of Southeast Asia. BMC Infectious Diseases, 20, 95-102.

Qiu, Y., et al. (2018). Machine Learning Approaches for Diagnosing Tuberculosis and Predicting Treatment Outcome. Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, 11, 43-49.

Sharma, S., & Singh, S. (2019). Machine Learning Approaches in Tuberculosis Diagnosis: A Review. Indian Journal of Medical Informatics, 24(4), 101-110.

Zhou, Y., et al. (2021). Evaluating the Use of Machine Learning in the Early Diagnosis of Tuberculosis: A Review. Journal of Clinical Medicine, 10(5), 1100