CASE DETECTION OF TUBERCULOSIS PATIENTS USING MACHINE-LEARNING

Patrick Kaggwa

2024-02-24

# 1. Abstract

## 1.1 Introduction

Although early detection and treatment of Tuberculosis cases are the hallmark of successful TB control, diagnostic delay is still long and common in Uganda. Therefore, the study aimed at developing a tool that uses machine-learning techniques to detect and diagnose TB more accurately in a shorter time.

## 1.2 Methods

This was a retrospective study that used secondary data collected between 2011 to 2018 collected from health under Infectious diseases research collaboration an (IDRC) project called MIND (Mulago Inpatient Non-invasive Diagnosis for pneumonia). The study was conducted on data from 2296 tested patients of whom 1345 had no TB and 951 had TB. Different machine learning metrics(Accuracy, RMSE and AUC) were used evaluate the performance of the models.

## 1.3 Results

Evaluation of four machine learning models—LASSO, Random Forest, Support Vector Machine (SVM), and Null model—revealed notable performance differences. Notably, the Random Forest model achieved the highest accuracy at 81.57% and an AUC of 0.8868, outperforming the other models.

## 1.4 Consclusion

The results highlight the potential of advanced machine learning algorithms, particularly Random Forest, in enhancing TB diagnosis and early detection strategies, offering valuable insights for clinical decision-making and the development of more effective screening protocols.

# 2. Introduction

Tuberculosis (TB) remains a significant global health challenge due to diagnostic limitations affecting both pediatric and adult populations, resulting in delayed diagnosis or misdiagnosis(Sagili et al., 2022). This delay impacts individual prognosis and community transmission. TB, caused by the bacterium Mycobacterium tuberculosis, primarily affects the lungs but can also involve other organs in its active form or remain latent without symptoms(Baykan et al., 2022),. In 2020, TB ranked among the top 10 causes of death globally, with over 10 million new cases and 1.3 million deaths reported. Africa bears a considerable burden, with an estimated quarter of new cases globally, resulting in approximately 417,000 deaths annually(Chakaya et al., 2022). TB is a leading cause of death among HIV-infected patients in sub-Saharan Africa, compounded by the low sensitivity of the commonly used sputum smear microscopy, particularly in detecting TB among people living with HIV(Palattiyil et al., 2022).

## 2.1 General Background Information

In Uganda, TB remains a significant public health concern, with an incidence of 330 cases per 100,000 people annually, including 136 new smear-positive cases per 100,000(Kintu et al., 2023). Diagnostic delays persist due to health services often waiting for systematic symptoms before examining sputum smears, leading to missed opportunities for timely diagnosis. The GeneXpert diagnostic tool offers superior sensitivity compared to sputum smear microscopy but is underutilized in Uganda due to its expense(Brown, Leavy, & Jancey, 2021), resulting in over 41,000 undiagnosed cases annually. Even with existing diagnostic methods like blood tests or sputum tests, analysis times are prolonged, allowing culture-positive TB cases to go undetected(Dong et al., 2022).

## 2.2 Description of data and data source

A thorough review of patients’ medical records was conducted to gather data on various risk factors associated with tuberculosis (TB). This involved examining all patient forms and their radiography reports, as well as analyzing laboratory test results obtained from the clinic’s records. The data collected included demographic variables such as gender and age, as well as clinical indicators such as oxygen consumption, asthma status, smoking habits, alcohol consumption, fever status, weight loss, cough status, sputum production, and presence of blood in sputum. Additionally, information on environmental factors such as the type of home fuel used was documented. Furthermore, HIV status and tuberculosis status were also recorded. This comprehensive approach to data collection ensured a thorough understanding of the factors influencing TB infection and progression within the patient population under study.

## 2.3 Questions/Hypotheses to be addressed

The prediction of TB can be modeled using early clinical patient data from Kampala residents who visited the IDRC between 2011 and 2018 and presented with a persistent cough lasting two weeks or more.

# 3. Methods

## 3.1 Study Design

The study was a retrospective study. This study was based on the secondary data, which was classified as clinical examination, patients’ history like alcohol use, smoking and diagnostics.

## 3.2 Study Population and Setting

The study population was the patients’ medical records between 2011 to 2018 who visited the hospital under an International Development Research Centre (IDRC) project. The project is currently running in China Friendship Hospital, Naguru.

## 3.3 Inclusion and Exclusion Criteria

The inclusion criteria was all patients who had cough for more than two weeks and the exclusion criteria was all adults living outside Kampala.

## 3.4 Data aquisition

The data was acquired from the TB clinic at Hospital patient records form patient forms and medical examination reports from the data records office. The TB dataset consisted of 2296 instances with 15 attributes.

## 3.5 Data import and cleaning

Our analysis began with thorough data preparation. Following a systematic approach, we ensured data integrity and reliability by sourcing and processing the dataset from an Excel file into R. To understand the data’s structure, we reviewed the codebook and conducted exploratory analyses, including summaries and visualizations. We then cleaned the data to enhance its quality for analysis. Character variables were converted to factors, and missing values were identified using visualizations and summaries and omitted. Finally, the data was strategically split into training and testing sets in an 70/30 ratio to prepare for model development and evaluation. These steps guaranteed a dataset ready for further analysis.

## 3.6 Model Development

Three models were considered for the prediction analysis, and each were fit for both outcomes of interest (tuberculosis). The data were split such that 70% were used for training the models and the remaining 30% were used for testing the final chosen model. Cross-validation was used for all models, using a ten-fold resampling structure. The tidymodel framework was utilized in the analysis.

## 3.7 Model Definition

There are several machine learning models suitable for addressing the research question, but this analysis primarily focused on three different types of models: Random Forest, LASSO, and Support Vector Machine (SVM).

Random Forest (RF): RF is a versatile and robust ensemble learning method that aims to reduce variance by building multiple decision trees on random subsets of the data and then averaging their predictions. It can handle large datasets with high-dimensional feature spaces and is less prone to overfitting compared to individual decision trees(Simon, Glaum, & Valdovinos, 2023). However, RF models are more complex and may be challenging to interpret compared to simpler models like decision trees.

LASSO: LASSO is a regularization-based linear regression method that balances model complexity and goodness of fit by penalizing the absolute values of the regression coefficients. This encourages sparsity in the model, meaning it tends to select a subset of the most important features while shrinking the coefficients of less important ones towards zero(Zhao & Yu, 2006). However, LASSO may struggle with highly correlated predictors and can only select one variable from a group of highly correlated variables, potentially ignoring important information.

Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm commonly used for classification tasks. It works by finding the optimal hyperplane that separates data points belonging to different classes with the maximum margin of separation. SVM is effective in high-dimensional spaces and is versatile in handling linear and non-linear relationships through different kernel functions.(JavaTpoint, 2021) However, SVM can be computationally intensive, especially with large datasets, and may require careful tuning of hyperparameters for optimal performance.

These three models were selected to provide a diverse set of approaches for predicting TB classification based on early clinical patient data. Random Forest offers a robust ensemble learning approach, LASSO provides a sparse and interpretable linear model, and SVM offers flexibility in handling complex relationships between predictors. By comparing and contrasting the performance of these models, we aim to identify the most effective approach for TB prediction in our study population.

## 3.8 Model Performance Evaluation

A confusion matrix or contingency table was used to determine the classification performance of the various classification models. Given m classes, the confusion matrix is a table of at least size m by m (Riehl, Neunteufel, & Hemberg, 2023). In a confusion matrix, if the instance is positive and it is classified as positive, it is counted as a true positive (TP); if it is classified as negative, it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as a true negative (TN); if it is classified as positive, it is counted as a false positive (FP). These terms are useful when analyzing a classifier’s ability to determine accuracy based on the correct and incorrect classes produced by the confusion matrix. Accuracy was used as it signifies the proportion of the total number of TB patient predictions that are correct(Carvalho, Pereira, & Cardoso, 2019). True positive means the number of actual TB patients that are accurately classified, and true negatives refer to the number of patients that are correctly classified as not having TB. The ROC (Receiver Operating Characteristic) curve was also used, which is a graphical representation to measure the performance of the classifier system. It shows the trade-off between sensitivity and specificity for every possible cut-off for a test(Robinson, Glen, & Lee, 2020). The area under the curve ranges from 0 to 1, where 1 implies an excellent test and 0 implies a useless test.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Data analysis was conducted using R Studio version RStudio RStudio 2024.04.0. Descriptive statistics (frequencies, mean, standard deviation (SD), median, interquartile range (IQR), and proportions) were used to present the baseline characteristics of participants. we used histograms, bar graphs, and box plots, to show how data was distributed.We did cross tabulation between the covariates and the outcome.A dataset of 2294 patients was analyzed. Amongst these patients, the gender distribution was 47.0% female and 53.0% male with mean age of 31 years.

## 4.2 Model results and accuracy assessment

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Model Perfomance for the Training Dataset   | Model | Accuracy | AUC | | --- | --- | --- | | Null Model | 0.5781931 | 0.5000000 | | Lasso Model | 0.7831776 | 0.8707883 | | Random Forest Model | 0.8255452 | 0.8973205 | | Support Vector model | 0.7838006 | 0.8618063 | |

This study examines the efficacy of four distinct machine learning models in predicting the presence of tuberculosis (TB) using input data. Each model, including LASSO, Random Forest, Support Vector Machine (SVM), and a Null model, is rigorously evaluated based on key performance metrics. The LASSO model applies the Least Absolute Shrinkage and Selection Operator regularization technique, while the Random Forest model employs ensemble learning. SVM seeks to optimize hyperplanes for class separation, and the Null model serves as a simplistic baseline. Performance assessment encompasses accuracy (with LASSO achieving 80.26%, Random Forest achieving 81.57%, SVM achieving 79.68%, and the Null model achieving 60.23%), Area Under the Curve (AUC) (with LASSO achieving 0.8895, Random Forest achieving 0.8868, SVM achieving 0.8825, and the Null model achieving 0.5). The findings underscore the potential of advanced machine learning algorithms, particularly Random Forest and LASSO, in enhancing TB diagnosis and early detection strategies.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Random Forest   | Prediction | Truth | Freq | | --- | --- | --- | | YES | YES | 633 | | NO | YES | 44 | | YES | NO | 152 | | NO | NO | 776 | |

The table presents the classification results of a Random Forest model for predicting tuberculosis (TB) status in individuals. Among individuals with TB (TB=YES), the model correctly identified 619 true positives and incorrectly predicted 219 cases as TB false negatives. For individuals without TB, the model accurately classified 709 cases as TB true negatives but misclassified 58 cases as TB false positives.

|  |
| --- |
| Figure 1: Random Forest ROC Curve |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Support Vector Machine   | Prediction | Truth | Freq | | --- | --- | --- | | YES | YES | 629 | | NO | YES | 48 | | YES | NO | 296 | | NO | NO | 632 | |

The table presents the classification results of a Support Vector Machine model for predicting tuberculosis (TB) status in individuals. Among individuals with TB, the model correctly identified 621 true positives and incorrectly predicted 291 cases as TB false negatives. For individuals without TB, the model accurately classified 637 cases as TB true negatives but misclassified 56 cases as TB positive (false positives).

|  |
| --- |
| Figure 2: Support Vector Machine ROC Curve |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4: lassoconf\_table   | Prediction | Truth | Freq | | --- | --- | --- | | YES | YES | 654 | | NO | YES | 23 | | YES | NO | 325 | | NO | NO | 603 | |

The table presents the classification results of a LASSSO model for predicting tuberculosis (TB) status in individuals. Among individuals with TB, the model correctly identified 654 true positives and incorrectly predicted 291 cases as TB false negatives. For individuals without TB, the model accurately classified 603 cases as TB true negatives but misclassified 23 cases as TB false positives.

|  |
| --- |
| Figure 3: LASSO ROC Curve |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 5: Model Perfomance Testing Results   | Model | Accuracy | AUC | | --- | --- | --- | | Null Model | 0.6023222 | 0.5000000 | | Lasso Model | 0.8026125 | 0.8894864 | | Random Forest Model | 0.8156749 | 0.8868349 | | Support Vector model | 0.7968070 | 0.8824730 | |

# 5. Models Evaluation

The evaluation of machine learning models’ performance on test data is pivotal for gauging their efficacy and generalization capabilities. The predictive accuracy of the models were measured using the test data. The Lasso, Random Forest, and Support Vector models were compared against a Null Model, with their respective performance metrics outlined. The results revealed a notable performance gap between the advanced models and the Null Model. Specifically, the Random Forest model exhibited superior performance, achieving an accuracy of 81.57% and an AUC of 0.887. In contrast, the Lasso and Support Vector models attained accuracies of 80.26% and 79.68%, with corresponding AUC values of 0.889 and 0.882, respectively.

# 6. Discussion

The study presents a comprehensive analysis of machine learning models’ efficacy in predicting the presence of tuberculosis (TB) based on a dataset comprising 2296 instances and 14 variables associated with TB. The study selected nine variables for prediction after conducting chi-square tests to eliminate non-statistically significant variables associated with the outcome. The retained variables for prediction included asthma, alcohol use, fever, chest radiography, home fuel use, weight loss, presence of blood in sputum, oxygen saturation, and HIV status. The study evaluated four distinct machine learning models, namely LASSO, Random Forest, Support Vector Machine (SVM), and a Null model, for their predictive performance using key metrics accuracy and Area Under the Curve (AUC). The LASSO model achieved an accuracy of 80.26% and an AUC of 0.8895. The Random Forest model outperformed others with an accuracy of 81.57% and an AUC of 0.8868. SVM achieved an accuracy of 79.68% and an AUC of 0.8825. The Null model achieved an accuracy of 60.23% and an AUC of 0.5.

Furthermore, the study presented detailed classification results for each model, emphasizing their ability to correctly classify true positives and true negatives while minimizing false positives and false negatives. Particularly, the Random Forest model demonstrated robust performance in correctly identifying true positives and true negatives, with minimal misclassifications. The findings underscore the potential of advanced machine learning algorithms, particularly Random Forest and LASSO, in enhancing TB diagnosis and early detection strategies. The manuscript provides valuable insights into the predictive capabilities of different machine learning models which can inform clinical decision-making and aid in the development of more effective TB screening tolls and protocols.

# 7. Strengths

In this study, a rigorous variable selection process was implemented, utilizing chi-square tests to identify statistically significant predictors associated with tuberculosis (TB) from a dataset comprising 2296 instances and 14 variables. By employing stringent selection criteria, the study ensures that only variables with meaningful associations with TB are included in subsequent analyses, thereby enhancing the model’s interpretability, and reducing the risk of overfitting. Moreover, the evaluation of multiple machine learning models, including LASSO, Random Forest, Support Vector Machine (SVM), and a Null model, contributes to a comprehensive comparison of different predictive approaches. This assessment allows for informed decision-making regarding model selection, facilitating the identification of the most suitable algorithm for TB prediction.

Additionally, the study employs robust performance metrics such as accuracy and Area Under the Curve (AUC) to evaluate model performance objectively. These metrics provide a comprehensive assessment of each model’s predictive capabilities, enabling comparisons across different algorithms. Furthermore, the detailed classification results presented offer insights into the models’ ability to correctly classify true positives and true negatives while minimizing false positives and false negatives.

# 8. Limitations and Recommendations

Despite the strengths, this study had limitations that warrant consideration. Firstly, the dataset’s potential class imbalance, characterized by variations in the number of instances for individuals with and without TB, may impact model performance of the model. Additionally, while the variable selection methodology employed chi-square tests, alternative feature selection methods like recursive feature elimination could offer further insights into the predictive variables’ relevance and importance. Future research should aim incorporate larger, more diverse datasets and conducting external validation studies to assess model performance in real-world clinical settings and use of more robust techniques such as neural networks.

# 9. Conclusions

The findings underscore the potential of machine learning algorithms, particularly Random Forest and LASSO, in enhancing TB diagnosis and early detection strategies. Future research should further advance the development of effective TB screening protocols and contribute to improved TB diagnosis and management on a global scale.

# 10. References

Baykan, A. H., Sayiner, H. S., Aydin, E., Koc, M., Inan, I., & Erturk, S. M. (2022). Extrapulmonary tuberculosıs: An old but resurgent problem. *Insights into Imaging*, *13*(1), undefined–undefined. <https://doi.org/10.1186/s13244-022-01172-0>

Brown, S., Leavy, J. E., & Jancey, J. (2021). Implementation of genexpert for tb testing in low- And middle-income countries: A systematic review. *Global Health Science and Practice*, *9*(3), 698–710. <https://doi.org/10.9745/GHSP-D-21-00121>

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. <https://doi.org/10.3390/electronics8080832>

Chakaya, J., Petersen, E., Nantanda, R., Mungai, B. N., Migliori, G. B., Amanullah, F., … Zumla, A. (2022). The WHO Global Tuberculosis 2021 Report – not so good news and turning the tide back to End TB. *International Journal of Infectious Diseases*, *124*, undefined–undefined. <https://doi.org/10.1016/j.ijid.2022.03.011>

Dong, B., He, Z., Li, Y., Xu, X., Wang, C., & Zeng, J. (2022). Improved Conventional and New Approaches in the Diagnosis of Tuberculosis. *Frontiers in Microbiology*, *13*, undefined–undefined. <https://doi.org/10.3389/fmicb.2022.924410>

JavaTpoint, U. (2021). Support Vector Machine Algorithm. *JavaTpoint*, undefined–undefined. Retrieved from <https://www.mendeley.com/catalogue/243af9d0-11e4-30d2-bdf5-97610716bdc8/>

Kintu, T. M., Mwanahamisi, B. S., Muwanguzi, M., Kyagambiddwa, T., Miiro, E., Tishekwa, N., … Nuwagira, E. (2023). Unfavorable treatment outcomes among patients with drug-resistant TB in Uganda. *International Journal of Tuberculosis and Lung Disease*, *27*(4), 291–297. <https://doi.org/10.5588/ijtld.22.0638>

Palattiyil, G., Kisaakye, P., Mwenyango, H., Katongole, S., Mulekya, F., Sidhva, D., … Bukuluki, P. (2022). Access to HIV/AIDS or TB care among refugees in Kampala, Uganda: Exploring the enablers and barriers during the COVID-19 pandemic. *Journal of Migration and Health*, *5*, undefined–undefined. <https://doi.org/10.1016/j.jmh.2022.100098>

Riehl, K., Neunteufel, M., & Hemberg, M. (2023). Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(5), 1394–1412. <https://doi.org/10.1093/jrsssc/qlad057>

Robinson, M. C., Glen, R. C., & Lee, A. A. (2020). Validating the validation: Reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *Journal of Computer-Aided Molecular Design*, *34*(7), 717–730. <https://doi.org/10.1007/s10822-019-00274-0>

Sagili, K. D., Muniyandi, M., Shringarpure, K., Singh, K., Kirubakaran, R., Rao, R., … Tharyan, P. (2022). Strategies to detect and manage latent tuberculosis infection among household contacts of pulmonary TB patients in high TB burden countries - a systematic review and meta-analysis. *Tropical Medicine and International Health*, *27*(10), 842–863. <https://doi.org/10.1111/tmi.13808>

Simon, S. M., Glaum, P., & Valdovinos, F. S. (2023). Interpreting random forest analysis of ecological models to move from prediction to explanation. *Scientific Reports*, *13*(1), undefined–undefined. <https://doi.org/10.1038/s41598-023-30313-8>

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563. Retrieved from <https://www.mendeley.com/catalogue/bb7b560c-7710-3e67-8975-46ef01e2d999/>