Prediction of Tuberculosis from Clinical Data

Patrick Kaggwa

2024-02-24

# 1. Abstract

#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 atdeveloping a tool that uses machine-learning techniques to detect and diagnose TB more accurately in a shorter time. #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. #Results

# 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(**Xin2019?**). This delay impacts individual prognosis and community transmission(F. Verbeke, G. Karara, & Van Bastelaere, 2012). 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(**Bulled2020?**),. 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(**WHO2023?**). 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.

## 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. 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, 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.

## 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

How can TB prediction be modeled using early clinical patient data?

What is the performance of the TB prediction models?

# 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.

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

Data analysis was conducted using R Studio version RStudio 2023.09.1. 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.

::: {#tbl-characteristics .cell tbl-cap=’ Participants' Characteristics’} ::: {.cell-output-display}

| X..Characteristic.. | X..NO….N…1.343 | X..YES….N…951 |
| --- | --- | --- |
| Gender | NA | NA |
| FEMALE | 646 (48%) | 443 (47%) |
| MALE | 697 (52%) | 508 (53%) |
| Asthma | NA | NA |
| NO | 1,284 (96%) | 934 (98%) |
| YES | 59 (4.4%) | 17 (1.8%) |
| Smoke | NA | NA |
| NO | 996 (74%) | 693 (73%) |
| YES | 347 (26%) | 258 (27%) |
| Alcohol | NA | NA |
| NO | 469 (35%) | 330 (35%) |
| YES | 874 (65%) | 621 (65%) |
| feverstatus | NA | NA |
| NO | 313 (23%) | 179 (19%) |
| YES | 1,030 (77%) | 772 (81%) |
| Weightloss | NA | NA |
| NO | 160 (12%) | 45 (4.7%) |
| YES | 1,183 (88%) | 906 (95%) |
| coughstatus | NA | NA |
| NO | 0 (0%) | 1 (0.1%) |
| YES | 1,343 (100%) | 950 (100%) |
| Sputumproduction | NA | NA |
| NO | 8 (0.6%) | 10 (1.1%) |
| YES | 1,335 (99%) | 941 (99%) |
| BloodinSputum | NA | NA |
| NO | 949 (71%) | 724 (76%) |
| YES | 394 (29%) | 227 (24%) |
| cxrconclusion\_chestxray | NA | NA |
| NEGATIVE | 860 (64%) | 333 (35%) |
| POSITIVE | 483 (36%) | 618 (65%) |
| homefuel | NA | NA |
| No | 852 (63%) | 15 (1.6%) |
| Yes | 491 (37%) | 936 (98%) |
| hivstatus | NA | NA |
| NEGATIVE | 530 (39%) | 303 (32%) |
| POSITIVE | 813 (61%) | 648 (68%) |

::: :::

##Determining requirements for Tuberculosis prediction that is appropriate data and modeling techniques

In our study, we utilized 2296 instances and examined 14 variables associated with TB. Subsequently, nine variables were selected for prediction after conducting chi-square tests. Non-statistically significant variables associated with TB were eliminated. Following this, chi-square tests were applied to the nine chosen variables to assess their association with a threshold of 0.05. However, these tests did not reveal a significant association among the variables themselves. Therefore, they were included in the prediction model. The variables retained for prediction were: asthma, alcohol use, fever, chest radiography, home fuel use, weight loss, presence of blood in sputum, oxygen saturation, and HIV status.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Model\_perfomance.   | 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.

::: {#tbl-Random Forest .cell tbl-cap=’ Model Predictions (Random Forest)’} ::: {.cell-output-display}

| Prediction | Truth | Freq |
| --- | --- | --- |
| YES | YES | 619 |
| NO | YES | 58 |
| YES | NO | 219 |
| NO | NO | 709 |

::: :::

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: rf\_ROCCURVE |

::: {.cell tbl-cap=’ Model Predictions (SVM)’} ::: {.cell-output-display}

| Prediction | Truth | Freq |
| --- | --- | --- |
| YES | YES | 621 |
| NO | YES | 56 |
| YES | NO | 291 |
| NO | NO | 637 |

::: :::

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: svm\_ROCCURVE |

::: {.cell tbl-cap=’ Model Predictions (LASSO)’} ::: {.cell-output-display}

| 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\_ROCCURVE |

#Figure1. Random Forest ROC Curve

#Models Evaluation

# 5. Discussion

The aim of the study was to apply machine-learning techniques in developing models for TB screening and detection in a shorter time based on patient data records to contribute to disease notification. The application of machine learning approaches have therefore been found to be successful in prediction (**Holzinger?**).

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References

F. Verbeke, G. Karara, & Van Bastelaere. (2012). *(4) (PDF) Patient Identification and Hospital Information Management Systems in sub-Saharan Africa: A Prospective Study in Rwanda and Burundi*. Retrieved from <https://www.researchgate.net/publication/269222217_Patient_Identification_and_Hospital_Information_Management_Systems_in_sub-Saharan_Africa_a_Prospective_Study_in_Rwanda_and_Burundi>