Investigating Machine Learning Methods for Tuberculosis Risk Factors Prediction – A Comparative Analysis and Evaluation

Oluwafemi Samson BALOGUN [0000-0003-2551-3059]
School of Computing, University of Eastern, Finland
samson.balogun@uef.fi

Sunday Adewale OLALEYE [0000-0002-0266-3989]
Department of Marketing, Management and International Business, University of Oulu, Finland
sunday.olaleye@oulu.fi

Mazhar MOHSIN
School of Computing, University of Eastern, Finland
mazhar.moshin@uef.fi

Pekka TOIVANEN
School of Computing, University of Eastern, Finland
pekka.toivanen@uef.fi

Corresponding Author: Oluwafemi Samson BALOGUN [0000-0003-2551-3059]

Abstract

Tuberculosis (TB) is a killer disease, and its root can be traced to Mycobacterium tuberculosis. As the world population increases, the burden of tuberculosis is growing along. Low-and-middle-income nations are not exempted from the tuberculosis crisis. Due to a shortage of medical supplies, tuberculosis bacteria have become a huge public health concern. This study reviewed recent literature from 2015 to 2020 to critically examine what earlier researchers have done about TB burden and treatment. The data used were based on the hospital's medical department's record and used a machine-learning algorithm to predict and determine the risk factors associated with the disease. Furthermore, it developed five predictive models to offer the medical managers a valid alternative to the manual estimation of TB patients' status as cured or not cured. The overall classification showed that all the classification methods performed well for classifying the TB treatment outcome (ranging between 67.5% and 73.4%). Our findings showed that MLP (testing) is the best model to predict TB patients' treatment outcomes. Age and length of stay were identified as significant risk factors for TB patients in this study. This study explains the study's limitation, contributions, managerial implications, and suggest future work.

Keywords: Tuberculosis. Prediction. Classification. Correlation. Machine learning

Introduction

Tuberculosis (TB) is a killer disease, and its root can be traced to Mycobacterium tuberculosis. According to Vivar et al. (2016), bacteria disease manifests in two forms. First, it could be Pulmonary, and this could account for 85% of cases. Second, it could be non-pulmonary, which accounts for 15% of cases. As the world population increases, the burden of tuberculosis is growing along. For example, World Health Organization (WHO) statistics nine years ago depict an increase in tuberculosis growth and estimate 7.8 million cases and 1.4 million deaths from tuberculosis (Ashna et al. 2018; Khaledi et al. 2016).

Low-and-middle-income nations are not exempted from the tuberculosis crisis. Due to a shortage of medical supplies, tuberculosis bacteria has become a huge public health concern. Statistics show that two million deaths are recorded yearly because of tuberculosis attacks (Harries and Dye 2006; World Health Organization 2006; Amiri et al. 2018). This crisis could be traced to tuberculosis patients' inability to complete their prescribed treatment (Cuneo and Snider 1989).

The extant assessment shows that failure to complete the treatment in industrialized nations is around 20% (Tangüis et al. 2000). The Center for Disease Control and Prevention in the United States supported the earlier proposition and confirmed that 25% of patients could not complete their chemotherapy (Yew 1999). Ultimately, the proportion of patients with active

Cite this Article as: Oluwafemi Samson BALOGUN, Sunday Adewale OLALEYE , Mazhar MOHSIN and Pekka TOIVANEN " Investigating Machine Learning Methods for Tuberculosis Risk Factors Prediction – A Comparative Analysis and Evaluation" Proceedings of the 37th International Business Information Management Association (IBIMA), ISBN: 978-0-9998551-6-4, 1-2 April 2021, Cordoba, Spain

disease who complete their therapy is between 20% and 40% in developing nations and 70%–75% in the USA (Legrand et al. 2008).

Noncompliance is a critical factor leading to tuberculosis's persistence in numerous nations, and the results of this well-recognized fact are prolonged infectiousness, relapse, prolonged and more expensive therapy due to multidrug-resistant TB, and death (Thiam et al. 2007). An earlier study revealed that noncompliance is related to a tenfold increase in poor outcomes from treatment and accounts for most treatment failures (Burman et al. 1997).

Noncompliance is the most significant issue hampering tuberculosis treatment and control: patients with active disease who are non-compliant have delayed sputum conversion to smear-negative, relapse rates five to six times higher, and potential of developing drug resistance (Legrand et al. 2008). That is, TB patients who stay in the pool of active cases will advance TB among latent cases who are inclined to be infected or affected.

Thus, TB threatens public health and creates enormous costs to improve public health improvement and advancement. Directly observed treatment, short-course (DOTS), a present worldwide control system for TB control, includes case detection and effective completion of the whole course of treatment. In 2006, to improve DOTS quality, the WHO designed a "Stop TB" Plan (World Health Organization 2006). In this plan, healthcare services ought to recognize and focus on interference factors that disrupt TB treatment. Supervision, conducted in a context-specific and patient-sensitive manner, assumes a significant role in patient treatment adherence and drug resistance prevention.

Although the WHO has highlighted the need to improve the nature of DOTS regarding supervision and patient support in the "Stop TB" plan, there is no particular method to quantify the necessary intensity of the health providers' support and supervision and to determine the TB cases in need of such supervision.

It is challenging to provide all TB patients with active supervision and support because of cost considerations and limited resources. Therefore, we may require an instrument to predict the patient's goal concerning TB treatment course completion. Such an instrument may distinguish TB patients at high-risk for treatment course noncompliance. This intervention may be applied to define the degree of supervision and support every patient need based on the predicted result of an exact predictive model. Currently, no framework is available to assess the TB treatment course by using TB patients' characteristics and designing a systematic method to predict the given result.

The defined result identified with each record of TB patients contains two potential classes: cured and not cured (desirable results) for the forecast of tuberculosis treatment. Ensuring that TB patients complete the treatment course is a primary step to TB control and public health promotion. The primary question of concern from previous studies is: "can new TB cases at risk of failing the treatment course completion be identified from early registration"? (Kalhori and Zeng 2013). For this reason, machine learning methods have already been applied, and they worked appropriately in the past investigations (Sitar-Taut 2009; Lazarescu 2002). An earlier study developed a model using historical datasets and assigned data to different pattern matching classes within the developed valid model (Serrano et al. 2006).

Some of the recent studies used secured SVM training over vertically partitioned datasets (Shen, Zhang, Zhu., Xu and Tang, (2019). Application of Naïve Bayes and Support Vector Machine algorithm (Li and Li, 2020), examination of hardware-efficient recognition accelerator with SVM classifier and cost-sensitive KNN classification (Zhang, 2020) while Sarkar, Vinay, Raj, Maiti, and Mitra (2019) used the application of optimized machine learning techniques for prediction of occupational accidents. This literature contributed to the body of knowledge with fewer machine learning models. The study of Shen et al. (2019) discovered a gap in their study and recommended a framework that supports multiple whiles (Li and Li, 2020), suggests classification algorithms comparison with more datasets. This study expanded the existing literature by using five distinct models to examine tuberculosis risk factors to fill the existing gaps with valid data.

To clarify the ambiguity in the existing literature, the objective of this work is to use a machine-learning algorithm to predict and determine the risk factors associated with tuberculosis disease and to develop five predictive models to offer the medical managers a valid alternative to the manual estimation of TB patients' status as cured or not cured.

This work starts with an introduction and reviews relevant literature conventionally. The third section explained the methodology utilized for this study and discussed the results—the study followed-up with contribution and conclusion. The work findings showed that MLP (testing) is the best model to predict TB patients' treatment outcomes.

Recent Literature on Tuberculosis

This study reviewed recent literature from 2015 to 2020 to critically examine what earlier researchers have done about TB burden and treatment. The relevant literature reviewed in this study shows different risk factors for TB. These risk factors could be shallow community knowledge, exposure to abattoir diseases, hesitant diagnosis, country of origin, incomplete treatment, and many more. The risk factors are country and regional-based. All the authors involved in the reviewed literature

have a common goal of seeking solutions to TB diseases. The management of existing and newly diagnosed TB was proposed towards international/regional adults and children. The different solutions proposed to need expert intervention. Besides, the literature shows implications that need health managers' attention. The managers need to consider the different seasons when thinking of intervention implementation and accelerate TB control programs to improve TB treatment. Most of the literature review proposed insightful future studies (see table 1 for more details).

Table 1: Literature review on treatment of TB

Author/ Date	Risk Factor(s)	Goals	Conclusions	Implications for Future research	Implications For practice
Asuquo et al. (2015)	Community knowledge	Reducing the burden of TB disease in 18 AKS societies through educational intervention and improving TB case detection and control.	The goal of the program to identify at least 300 TB cases reported by the public-private partnership has been achieved despite many challenges.	The program has saved between 3000 and 3600 susceptible individuals from being infected. The authors hoped that the efforts made by this program would be sustained.	An undetected case has the potential of infecting 10–12 susceptible individuals annually.
Sa'idu et al. (2017)	Disease contracted through abattoir had a seasonality trend as a risk factor that confounded the occurrence of the disease in the Gombe region	To study retrospective abattoir bovine tuberculosis in Gombe Township in Northeastern Nigeria from 2008 to 2015 using the abattoir records.	The research concluded that bovine TB is prevalent in Northeastern Nigeria, indicating epidemic status in Gombe state.	Future studies need to examine seasonal factors that have a significant effect on bovine TB in the north-eastern region of Nigeria.	The distribution and occurrence of the bovine TB disease in the rainy season were higher than in the dry season.
Swain et al. (2019)	Delayed diagnosis or untreatable laryngeal tuberculosis	To assess the clinical appearance, diagnosis, and treatment of primary laryngeal tuberculosis at a tertiary care hospital in Eastern India.	This study concluded that delayed diagnosis or untreatable laryngeal tuberculosis leads to high morbidity and mortality of the patient, and it is crucial to have a high index of suspiciousness to rule out tubercular lesion in the larynx as this disease is curable.	Primary laryngeal tuberculosis is a highly contagious human disease. Laryngeal tuberculosis has nonspecific clinical presentations, and future research is necessary on primary laryngeal tuberculosis.	Larynx tuberculosis is frequently confused with syphilis, fungal laryngitis, and granulomatous lesions such as Wegner granulomatosis and sarcoidosis. It is crucial to explain the existence of tuberculosis.
Langer et al. (2019)	Country of birth or country of the long history of residence.	To formulate measures to prevent TB transmission in the United States to avoid potential increases in recent transmissions that could lead to large outbreaks.	Proposed measures to prevent TB transmission in the United States.	Future US TB prevention efforts should include a focus on testing for and treating latent tuberculosis infection to prevent progression to tuberculosis disease.	As the US TB control program enters a new decade, a combination of old and new approaches is needed to maintain and accelerate eliminating TB in the United States. More emphasis is required on testing for and treating LTBI in high-risk populations, such as the non–US-born.
Katiyar and Katiyar (2019)	Socio-economic conditions	To improve case finding and effectively treat patients of tuberculosis both in the	Provided protocol for the management of newly diagnosed cases of tuberculosis.	There are many recommendations available for the management of tuberculosis, which are considered very	Improving the treatment of tuberculosis in the private sector, and ensuring the quality of care, a protocol appropriate to our socio-economic conditions is

		public and private sector.		useful starting points for care. However, these are not the only therapeutic choices, as the recommendations cannot cover every potential scenario and replace sound professional judgment.	required. Such a protocol can provide guidelines for improved care outcomes and help to disrupt the spread of the disease in the community and minimize drug resistance growth.
Khatami et al. (2019).	Children rapid disease progression	To probe into how to manage children with tuberculosis	In both cases, microbiological confirmation should be obtained. Even the children should be treated according to the most likely source event's drug susceptibility pattern in the absence of microbial proof.	Tuberculosis in young children can present with acute or non-specific signs and symptoms.	The availability of child-friendly formulations remains an issue.
Akkerman et al. (2019).	Comprehensiv e surveillance	To demonstrate the feasibility of introducing national aDSM registers and identifying the form and frequency of adverse events associated with exposure to new anti-TB drugs.	The measures taken resulted in 100% coverage for most countries surveyed, although the entire range was lower in some of them.	Monitoring of adverse events in the treatment of drug-resistant tuberculosis is difficult to accomplish in many countries.	The recruitment process on all continents was lengthy and time-consuming, although the support and enthusiasm of participating colleagues allowed for resolving any existing problems. Several countries (including sub-Saharan Africa) were invited to participate. However, some centers decided to refuse because the initiative was voluntary. And the operation was considered "difficult" or "time-consuming" without additional funding.
Levillain et al. (2019)	Strains of the Beijing lineage	To make a preclinical assessment of a new live attenuated Mycobacteriu m tuberculosis.	This study highlighted the differences in protection efficacy of live attenuated M. tuberculosis-derived vaccine candidates depending on their genetic background and provided insights for the development of novel live vaccines against TB, especially in East-Asian countries where M. tuberculosis strains of the Beijing family are highly dominant.	A vaccine that is better than BCG is urgently needed.	One obstacle for new TB vaccines is to protect against all Mycobacterium tuberculosis lines, including the most virulent ones, such as the Beijing lineage.
Hosseini et al. (2020)	Prolonged treatment with antibiotics or immunosuppre ssive agents	To conduct a review of cross-sectional studies on the prevalence of	Aspergillus-coinfection is a common occurrence in tuberculosis patients; thus, patients	This study could not conduct subgroup analysis and meta-regression due to insufficient data on	The present analysis found a robust combined Aspergillus coinfection amongst patients with pulmonary tuberculosis in Asia and Africa.

		pulmonary Aspergillus coinfection among patients with pulmonary tuberculosis according to the PRISMA Protocol	should be closely monitored for Aspergillus's problems.	alcohol consumption, BMI, participants (adults, children, and neonates), smoking, DM, etc. These factors may significantly affect the variability of the estimates of prevalence, and this remains a gap for future study.	
Lu et al. (2020)	The non-adherence of patients to the lengthy treatment, adverse effects of the drugs, and the emergence of multi-drug resistant strains.	To examine the anti-mycobacterial activity of a plant-derived triterpenoid, sophoradiol, against the drug-resistant strains of Mycobacteriu m tuberculosis and in the murine model of tuberculosis.	Cytotoxicity assays have shown that sophoradiol has negligible toxicity to regular human breast cell lines. The study concluded that sophoradiol might prove to be a beneficial lead molecule for the management of tuberculosis.	The cytotoxicity of molecules should always be considered as the drug may exhibit significant adverse effects on the human body.	The results showed that sophoradiol exhibited unusual behavior against the H37RV strain with an 8.5 µg / mL MIC.

Materials and Method

This study utilized secondary data from the Department of Medical records of the University of Ilorin Teaching Hospital, Nigeria. The data is on patients treated for TB and spans fifteen years. The data collected included the patient's age, sex, length of stay, and treatment outcomes (cured and not cured). This study performed the analysis using IBM SPSS software version 25.

Correlation Analysis

When examining a large dataset, the data's insight is not valuable unless the noise is removed because an irrelevant feature only occupies space and does not add anything meaningful and new to the target phenomenon (Guyon and Elissee, 2003). To ascertain the predictors, influence on target variables, this study conducts a bivariate correlation. This data analysis technique allows the relationship between the predictors and the response variables. The study uses this technique to calculate the correlation coefficient "r" as proposed by (Dash and Liu,1997). Due to uncertain prediction, the study employed a two-tailed test hypothesis. For instance, if we have two variables, X as a list of features and Y as the class, the optimal subset is always relative to the evaluation function. The intensity of variable relatedness determines according to Field's (2015) correlation criteria. It indicates that the linear relationship between variables ranges from +1 to −1. Positive linear relationship precision among variables could be +1, while -1 indicates the opposite of positive linear association.

The common formula for the calculation of correlation coefficient “r” is

$$r = \frac{n(\sum XY) - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}},$$

taking that X and Y variables with n representing TB instances, it is possible to use t test for correlation coefficient to determine the statistical significance. The below formula is applicable

$$t = \sqrt{\frac{(n-2)}{(1-r^2)}}$$

The t table can used for variables significance relationship i.e. X and Y ($p < 0.05$)

degrees of freedom for correlation coefficient calculation is equal to $n - 2$. Then a t-table is used to find a significant relationship between each of the variables X and Y ($P < 0.05$).

Problem Modeling

Several classification algorithms that are meant for TB treatment were employed to reach the goal of this study. Han and Kamber (2016) remarked that 'prediction can be viewed as the construction and use of a model to assess an unlabeled sample class or evaluate the value or value ranges of an attribute that a given sample is likely to have.' Based on our comparison, this study adopts five classifiers, namely binary logistic regression (BLR), discriminant analysis (DA), multilayer perceptron (MLP), radial basis perceptron (RBF), and decision tree (DT). The study expatiates on the relevance of selected classifiers in the next section. The study applies five classifiers to train the dataset to estimate the relationship among the attributes and build predictive models. The model with high precision will be selected to predict the outcome of TB treatment.

Radial Basis Function (RBF)

An artificial intelligence network that uses a radial basis in a linear combination as an activation function is a radial basis function (RBF) network. This type of network was designed for viewing a problem in curve-fitting (approximation) and high-dimensional space. According to Marsland (2009), the RBF technique's motivation is to find a multidimensional feature that provides the best fit for a training tuple and subsequently applies this multidimensional surface to interpolate test data via regularization. The Gaussian radial base function is used in this study. The prominent RBF architecture consists of three layers: input, middle (hidden), and output. The input layer is provided with the requisite information fed into the network. Hidden layers specify the weights between the input and the hidden unit to trigger the hidden unit. The inputs will decide the roles of these layers. The output layer's function depends on the hidden unit's activity and the weight of the correlation between the hidden output unit.

Artificial Neural Network

Artificial neural networks (ANNs) are biologically based computational approaches that can model complex non-linear functions. A standard architecture called a multi-layer perceptron (MLP) with a back-propagation algorithm is developed. The neural network is a compound of connected input/output units in which each link has an associated weight. Adjusting weights is a crucial step in predicting the suitable class label input through iterative learning. This approach is commonly used in classification and prediction tasks due to its high noise tolerance and the classification of unseen patterns (Vittinghoff et al., 2011).

Decision Tree

A nonparametric estimation algorithm where the input space is divided into local regions defined by distance measurement, such as the Euclidean standard, is known as a decision tree. It is a flowchart-like tree structure where the inner node, branches, and leaf node are concepts associated with training tuples. In this hierarchical data structure, the local area is defined in a sequence of recursive splits in a limited number of steps by applying a divide-and-conquer strategy. This robust classifier is famous for its intuitive explanation (Vittinghoff, 2011). The chi-square automatic interaction detection technique is used in this research work.

Binary Logistic Regression

Logistic regression (LR) is used principally for predicting binary or multi-class dependent variables. This algorithm's response variable is discrete, and it builds a model to predict the odds of its occurrence. The limiting assumptions of normality and independence of this method have contributed to an increase in the application and popularity of machine learning techniques to real-world prediction problems. In this analysis, binary logistic regression, an algorithm that constructs a hyperplane separating two data sets using a logistic function to express distance from the hyperplane as a likelihood of dichotomous class membership, was used:

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\cdots+\beta_n X_n+\varepsilon)}} \quad (1)$$

This equation X_i symbolizes discrete or continuous predictor variables with numeric values. In the case of a dependent variable (Y) being dichotomous, we use this algorithm. The constants $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients estimated from the training data, typically computed using an iterative maximum likelihood technique. Usually, this of formula justification is that the log of the odds, a number that goes from $-\infty$ to $+\infty$, is a linear function. Mainly by using this

model, stepwise selection of the variables is applied, and the related coefficients are calculated. In producing the LR equation, the variables' statistical significance is determined by the maximum-likelihood ratio (Alpaydin, 2014).

Discriminant Analysis

The elements of the discriminant models are given as $Z = a + W_1X_1 + W_2X_2 + ... + W_kX_k$,

where Z is the discriminant score, a is the discriminant constant, W_k is the discriminant weight or coefficient, and X_k is an independent or predictor variable. Discriminant analysis is a multivariable technique that separates distinct sets of observations and attributes new observations to predefined sets. The statistical problem is to develop a law (discriminant function) based on population size. According to this law, new samples with no clear attribution are attributed to one of the populations. Fisher discriminant analysis can be pointed out as one of the most known functions applied in discriminant analysis.

Result and Analysis

Descriptive and Correlation

Table 2 contains the demographic profiles of the TB patients. The TB patients' demographic breakdown shows that most of the patients are within age 26-35 and the average age is 38 years, majority of the patients are male, and the majority of the patients stay within 15-25 months while their average length of stay is 20 months.

Table 2: Demographic profiles of the TB patients

Attributes	Scale	Frequency	Percentage	Average
Age	Below 15	55	0.2%	38years
	15 – 25	87	10.6%	
	26 – 35	127	24.5%	
	36 – 45	98	18.9%	
	Above 45	150	29.9%	
Gender	Male	298	57.5%	
	Female	220	42.5%	
Length of Stay	Below 15	248	47.9%	20months
	15 – 25	134	25.9%	
	26 – 35	55	10.8%	
	36 – 45	39	7.5%	
	Above 45	41	7.9%	

Table 3 shows the correlation among the various selected feature attributes. Generally, there is no significant linear relationship between the feature attributes except significant ones. However, the correlation among the individual attributes varied. For instance, while there exists weak negative correlation between age and length of stay ($r = -0.103$), there is a positive weak correlation between gender and length of stay ($r = 0.021$) while there is a weak positive correlation between age and gender ($r = 0.071$).

Table 3: Correlation among the various feature attributes

	Age			Gender			Length of stay		
	Pearson	P-value	Remark	Pearson	P-value	Remark	Pearson	P-value	Remark
Age				0.071	0.106	Not significant	-0.103	0.019	Significant
Gender							0.021	0.631	Not significant

Classification Algorithms Analysis

This section shows the results and figures from the different classifiers used for the TB patients.

Table 4: Model Summary

The table below shows the variability explained of the outcome variable explained by the logistic model.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	618.649	0.112	0.153

Table 5: Hosmer and Lemeshow Test

The table below indicates how well the model fits the data used.

Step	Chi-square	Degree of freedom	Sig.
1	13.647	8	0.091

Table 6: Eigenvalues

This table shows the relationship between the discriminant function and group variable.

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.115	100	100	0.322

Table 7: Wilks' Lambda

This shows the contributions of all the independent variables used

Test of Function(s)	Wilks' Lambda	Chi-square	Degree of freedom	Sig.
1	0.897	56.032	3	0.001

Table 8: Variables and Coefficients for Discriminant Analysis, Logistic Regression, and Decision Tree

This table shows the individual contributions of independent variables used for the model.

Methods	DA			BLR			DT	
Independent Variable	Wilks' Lambda	Canonical Coefficient	Sig.	Wald Statistic	Coefficient	Sig.	Chi-Square	Sig.
Constant				9.832	0.866			
Age	0.948	0.646	0.001	20.189	-0.024	0.001	8.488	0.001
Length of stay	0.939	-0.723	0.001	25.755	0.037	0.001	73.781	0.033
Gender	0.998	0.098	0.372	0.409	-0.126	0.523		

Table 9: Variable Importance for Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF)

This table shows the order of importance of the independent variables used.

Methods	MLP		RBF	
Independent Variable	Importance	Normalized Importance	Importance	Normalized Importance
Age	0.495	100%	0.423	99.2%
Length of stay	0.454	91.1%	0.427	100%
Gender	0.055	10.9%	0.150	35.2%

Table 10: Classification of Tuberculosis Outcome Using Discriminant Analysis (DA), Logistic Regression, and Artificial Neural Network

This table shows the overall percentage correctly classified for each of the method used.

	Predicted Group Membership		
Model	Actual Group	0	1
DA	0	61(32.1%)	129(67.9%)
	1	39(11.9%)	288(88.1%)
Overall % Correctly Classified	**67.5%**		
BLR	0	70(36.8%)	120(63.2%)
	1	46(14.1%)	281(85.9%)
Overall % Correctly Classified	**67.9%**		
MLP(Training)	0	55(44.7%)	68(55.3%)
	1	27(12.2%)	194(87.8%)
Overall % Correctly Classified	**72.4%**		
MLP(Testing)	0	31(46.3%)	36(53.7%)
	1	10(9.4%)	96(90.6%)
Overall % Correctly Classified	**73.4%**		
RBF(Training)	0	63(47.7%)	70(52.6%)
	1	40(17%)	195(83%)
Overall % Correctly Classified	**70.1%**		
RBF(Testing)	0	26(45.6%)	31(54.4%)
	1	24(26.1%)	68(73.9%)
Overall % Correctly Classified	**63.1%**		
DT	0	72(37.7%)	119(62.3%)
	1	33(10.1%)	294(89.9%)
Overall % Correctly Classified	**70.7%**		

Table 11: The Value of the Performance Criteria for BLR, DA, MLP, and RBF in Predicting the TB Treatment Outcome

The table below shows how effective the methods used predict the treatment outcome of the TB patients.

Methods	Set	Sensitivity	Specificity	Positive Predicted Value	Negative Predicted Value
BLR		61.2%	70.1%	36.8%	85.7%
DA		60.3%	70.1%	36.8%	85.9%
MLP	Training	67.1%	74%	44.7%	87.8%
	Testing	75.6%	72.7%	46.3%	90.6%
RBF	Training	61.2%	73.6%	47.4%	83.0%
	Testing	52.0%	68.7%	45.6%	73.9%
DT		68.8%	71.2%	37.7%	89.9%

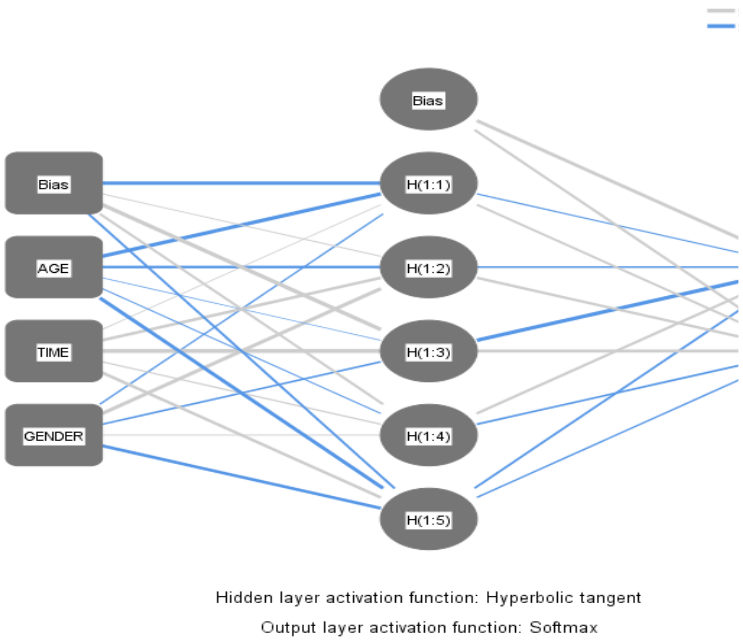

Fig. 1: Architecture of the multilayer perceptron

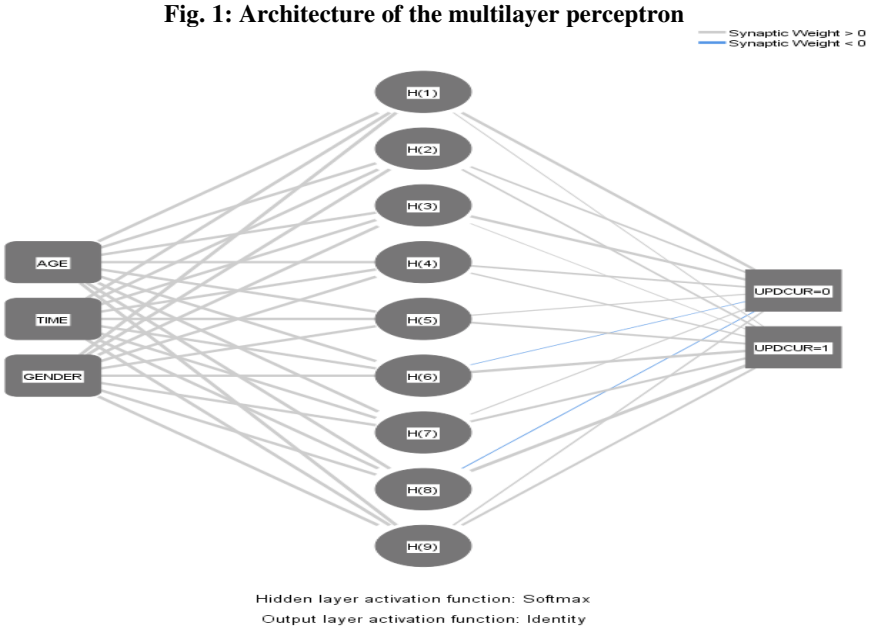

Fig .2: Architecture of the radial basis function

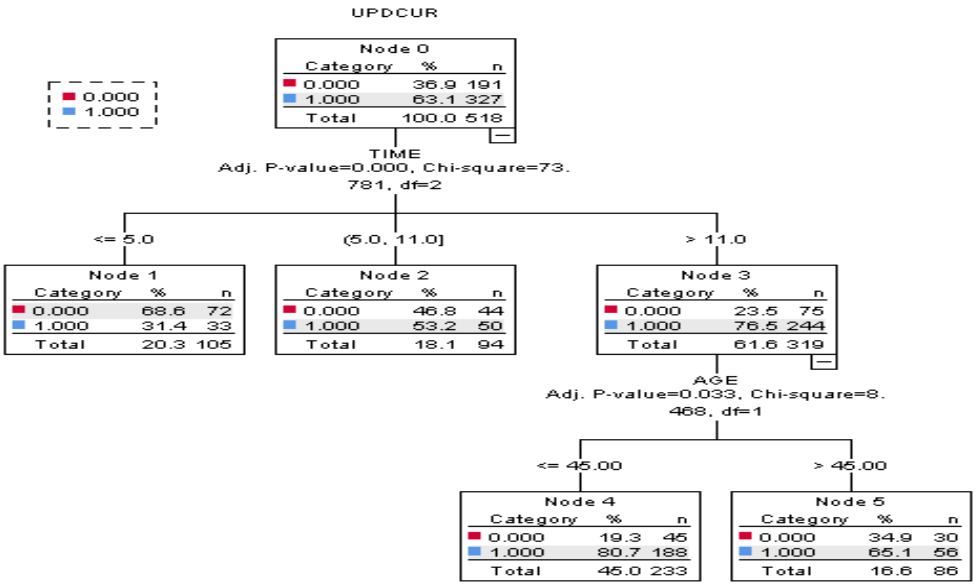

Fig. 3:Flowchart of the decision tree

Discussion

This study applied some machine learning methods to determine predicted treatment outcomes and risk factors associated with TB patients. In this regard, five machine learning models, BLR, DA, MLP, RBF, and DT, were used.

The overall classification showed that all the classification methods performed well for classifying the TB treatment outcome (ranging between 67.5% and 73.4%), See table 10. In other words, all classification methods were quite efficient in predicting the TB treatment outcome. In terms of sensitivity, the training and testing of the MLP predicted the treatment outcome of the patients well. As the minimum sensitivity was 52% (testing of RBF), and the maximum value was 75.6% (testing of MLP). Besides, the methods' performance in terms of specificity, overall classification, and sensitivity for predicting TB treatment outcome was good (ranging between 52% and 75.6%). All the methods had a sensitivity more significant than 50% (See table 11). However, in previous studies, it was reported that none of the methods have sensitivity more significant than 50% (Tapak et al., 2019).

Considering that the treatment outcome is a crucial prediction in biomedical applications, a higher sensitivity classifier is preferred. Therefore, an efficient classification method must have the ability to predict the potential treatment outcome of patients using the predictor variables. As shown, all the classifiers performed well enough to predict the patients' treatment outcome in our study.

The Nagelkerke R square indicates that 15.3% of the total variation in the outcome variable (cured and not cured) explained by the logistic regression model fitted into the data (see table 4). The Hosmer and Lemeshow test are not significant, which suggests that the model fit the data well ($p > 0.05$) (see table 5). The eigenvalue shows the linear combination of only one with eigenvalue (0.115), and 100% of the variance accounted for the group variable (cured and not cured). Besides, the canonical correlation (0.322) shows a lesser relationship between the discriminant function and the group variable (see table 6).

Table 7 shows that the Wilks' Lambda was significant, suggesting a relationship between the discriminant function and group variable ($p < 0.05$). Wilks' lambda was used to test which independent variables contribute significantly to the discriminant function. The p-value shows that two of the independent variables: age and length of stay, are highly significant risk factors ($p < 0.05$), while gender is not a significant risk factor ($p > 0.05$) (See table 8).

For the logistic regression, the classification coefficient was used to assess the relative classifying importance of the dependent variable (cured and not cured). The Wald statistic from the logistic regression analysis was used to test the null hypothesis that the model's independent variables' coefficients are zero. It shows that two variables, age, and length of stay, are significant risk factors at ($p < 0.05$), while gender is not a significant risk factor at ($p > 0.05$) see table 8.

Using the chi-square, DT showed that age and length of stay were significant risk factors ($p > 0.05$). Besides, MLP and RBF were used to determine the relative classifying importance of the dependent variables (cured/not cured), and they showed that age and length of stay were essential risk factors, while gender was not. Overall, the five methods identified the same variables: age and length of stay as risk factors and gender, not risk factors (see Tables 8 and 9).

The discriminant analysis's overall accuracy in predicting the TB treatment outcome (cured and not cured) was 67.5% (See table 10). The discriminant model is: (See table 9). The logistic regression model's overall accuracy in predicting the TB treatment outcome with a predicted probability of 0.5 or greater is 67.9%. The result shows no problem of over-fitting because the overall classification of training and testing is very close for MLP and RBF (See table 10). Figures 1-2 show the architectural design of TB patients' classification for MLP and RBF, while Figure 3 shows the flowchart of TB patients' classification for DT.

Many studies have compared the performance of various classification methods to predict the outcome of interest. This study focused on the performance of different classification methods in predicting and determining the risk factors associated with TB patients' treatment outcomes. Our findings showed that MLP (testing) outperformed the other machine learning methods in predicting patients' treatment outcomes in terms of some criteria. Besides, their performance in terms of sensitivity was good for the prediction. The machine learning methods utilized in this study are nonparametric, and they provide efficient solutions for classification problems without considering any unique assumptions regarding the distribution of data. They also deal with nonlinearity and high-order interactions (Alpaydin, 2014). However, the performance of a method is dependent on data, and in general, no method always performs as the best technique in classification problems.

The classifiers identified age and length of stay as significant risk factors, while gender was not a significant risk factor for TB patients.

Contribution

Our result is similar to the study of Bourdès et al. (2010); in their study, the Neural Network technique outperformed others compared to Logistic Regression, consistent with this study. Nevertheless, the study compares two techniques, but our study compared four machine learning techniques. Our result is contrary to the study of Mansour, Eghbal & Amirhossein (2013), in their Logistic Regression is the best method in comparison to Discriminant analysis and RBF, which is not consistent with this study. However, the study compares three machine learning techniques while ours compares four techniques. Also, our study's result is like Goss & Ramchandani's (1995) study, showing that Neural Network performs better than others compared to Binary Logistic Regression and Discriminant analysis is consistent with this study. Nonetheless, the authors compare three methods of machine learning techniques, but our study compared four techniques.
Knowing the best method to ascertain the risk factors associated with TB is a herculean task due to TB disease's complexity. Also, knowing the best practice in using a machine learning approach is not easy. The study shows the best machine learning techniques with the best predictive features and contributes to machine learning literature by integrating four machine learning techniques. Moreover, the result shows that MLP (testing) gives the highest classification rate in predicting patients' treatment outcomes.

Managerial Implication

This study will help health managers and researchers to predict and identify the risk factors associated with TB disease. Also, methodological-wise, this will help the practicing managers to know how to make the right decisions when confronted with a different methodological approach.

Conclusion

The data used for this study was collected from the hospital's medical department's record, making the analysis vulnerable to possible biases for the estimation of criteria such as sensitivity, which is a limitation for this study. This study focused on evaluating the performance of five machine learning methods in predicting treatment outcomes and determining the risk factors associated with TB patients. Our findings showed that MLP (testing) is the best model to predict TB patients' treatment outcomes. Age and length of stay were identified as significant risk factors for TB patients in this study. Further investigation is needed using large datasets and more factors to recommend a helpful treatment outcome prediction tool. Based on our results, we recommend that health professionals and researchers adopt MLP for future prediction of infectious disease treatment outcomes. Lastly, further research will need to consider other factors to identify the risk factors and developing a robust epidemiological model. It is essential to consider other factors.

Conflicts of Interest

We declare no conflicts of interest.

References

- Akkerman, O., Aleksa, A., Alffenaar, J. W., Al-Marzouqi, N. H., Arias-Guillén, M., Belilovski, E. and Loidi, J. C. (2019), 'Surveillance of adverse events in the treatment of drug-resistant tuberculosis: A global feasibility study', *Int J Infect Dis*, 83, 72-76.
- Alpaydin, E. (2014), Introduction to Machine Learning, 3rd Edition. MIT Press, Cambridge.
- Amiri, M. R. J., Siami, R. and Khaledi, A. (2018), 'Tuberculosis status and coinfection of pulmonary fungal infections in patients referred to reference laboratory of health centers Ghaemshahr City during 2007–2017', *Ethiop J Health Sci,* 28(6), 683-690. https://doi.org/10.4314/ejhs.v28i6.2.
- Ashna, H., Kaffash, A., Khaledi, A. and Ghazvini, K. (2018), 'Mutations of rpob gene associated with rifampin resistance among Mycobacterium tuberculosis isolated in tuberculosis regional reference laboratory in northeast of Iran during 2015–2016', *Ethiop J Health Sci,* 28(3), 299-304.
- Asuquo, A. E., Pokam, B. D. T., Ibeneme, E., Ekpereonne, E., Obot, V. and Asuquo, P. N. (2015), 'A public–private partnership to reduce tuberculosis burden in Akwa Ibom State, Nigeria', *Int J Mycobacteriol,* 4(2),143-150.
- Bourdès, V., Bonnevay, S., Lisboa, P., Defrance, R., Pérol, D., Chabaud, S., Bachelot, T., Gargi, T. and Négrier, S. (2010), 'Comparison of artificial neural network with logistic regression as classification models for variable selection for prediction of breast cancer patient outcomes', *Advances in Artificial Neural Systems*.
- Burman, W. J., Cohn, D. L., Rietmeijer, C. A., Judson, F. N., Reves, R. R. and Sbarbaro, J. A. (1997), 'Noncompliance with directly observed therapy for tuberculosis: epidemiology and effect on the outcome of treatment', *Chest*, 111(5), 1168–1173.
- Cuneo, W. D. and Snider, D. J. (1989), 'Enhancing Patient Compliance with Tuberculosis Therapy', *Clin Chest Med*, 10(3), 375-380.
- Dash, M. and Liu, H. (1997), 'Feature selection for classification', *Intell Data Anal*, 1(3), 131-156.
- Field, A. (2015), Discovering Statistics using SPSS. 2nd Edition, SAGE Publication LTD, London.
- Goss, E. P. and Ramchandani, H. (1995), 'Comparing classification accuracy of neural networks, binary logit regression and discriminant analysis for insolvency prediction of life insurers', *Journal of Economics and Finance*, 19(3), 1.
- Guyon, I. and Elissee, A. (2003), 'An introduction to variable and feature selection', *J Mach Learn Res,* 3(3), 1157-1182.
- Han, J. and Kamber, M. (2006), Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, Burlington.
- Harries, A. D. and Dye, C. (2006), 'Tuberculosis', *Ann Trop Med Parasitol*, 100(5), 415-443. https://doi.org/10.1179/136485906X91477.
- Hosseini, M., Shakerimoghaddam, A., Ghazalibina, M. and Khaledi, A. (2020), 'Aspergillus coinfection among patients with pulmonary tuberculosis in Asia and Africa countries: A systematic review and meta-analysis of cross-sectional studies', *Microb Pathog,* 104018.
- Kalhori, S. R. and Zeng, X. J. (2013), 'Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course', *J Intell Learn Syst Appl,* 5, 184-193.
- Katiyar, S. K. and Katiyar, S. (2019), 'Protocol for the management of newly diagnosed cases of tuberculosis', *Indian J Tuberc,* 66(4), 507-515.
- Khaledi, A., Bahador, A., Esmaeili, D., Tafazoli, A., Ghazvini, K. and Mansury, D. (2016), 'Prevalence of nontuberculous mycobacteria isolated from environmental samples in Iran: a meta-analysis', *J Res Med Sci: The Official Journal of Isfahan University of Medical Sciences*, 21,58, https://dx.doi.org/10.4103%2F1735-1995.187306 .
- Khatami, A., Britton, P. N. and Marais, B. J. (2019), 'Management of Children with Tuberculosis', *Clin Chest Med*, 40(4), 797-810.
- Langer, A. J., Navin, T. R., Winston, C. A. and LoBue, P. (2019), 'Epidemiology of Tuberculosis in the United States', *Clinics in Chest Medicine,* 40(2), 693-702.
- Lazarescu, M., Turpin, A. and Venkatesh, S. (2002), 'An application of machine learning techniques for the classification of glaucomatous progression', In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR) (pp. 243-251). Springer, Berlin, Heidelberg.
- Legrand, J., Sanchez, A., Le Pont, F., Camacho, L. and Larouze, B. (2008), 'Modeling the impact of tuberculosis control strategies in highly endemic overcrowded prisons', *PLoS One,* 3(5), 1-10.
- Levillain, F., Kim, H., Kwon, K. W., Clark, S., Cia, F., Malaga, W. and Bancroft, G. J. (2019), 'Preclinical assessment of a new live attenuated Mycobacterium tuberculosis Beijing-based vaccine for tuberculosis', *Vaccine*, 38(6), 1416-1423.

- Li, Q. N., & Li, T. H. (2020). Research on the application of Naive Bayes and Support Vector Machine algorithm on exercises Classification. In Journal of Physics: Conference Series (Vol. 1437, No. 1, p. 012071). IOP Publishing.
- Lu, N., Yang, Y., Liu, J., Li, J., Ouyang, B., Xia, J. and Du, Y. (2020), 'Sophoradiol inhibits the growth of drug resistant Mycobacterium tuberculosis in vitro and murine models of tuberculosis', *Microb Pathog*, 103971.
- Luo, A., An, F., Zhang, X., & Mattausch, H. J. (2019). A hardware-efficient recognition accelerator using Haar-like feature and SVM classifier. IEEE Access, 7, 14472-14487.
- Mansour, R., Eghbal, K. and Amirhossein, H. (2013), 'Comparison of artificial neural network, logistic regression and discriminant analysis efficiency in determining risk factors of type 2 diabetes'.
- Marsland, S. (2009), Machine Learning: An Algorithmic Perspective. 1st Edition, Chapman and Hall, London.
- Sa'idu, A. S., Mohammed, S., Ashafa, M., Gashua, M. M., Mahre, M. B. and Maigado, A. I. (2017), 'Retrospective study of bovine tuberculosis in Gombe township abattoir, Northeastern Nigeria', *Int J Vet Sci Med*, 5(1), 65–69.
- Sarkar, S., Vinay, S., Raj, R., Maiti, J., & Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research, 106, 210-224.
- Serrano, J. I., Tomeckova, M. and Zvárová, J. (2006), 'Machine learning methods for knowledge discovery in medical data on Atherosclerosis', *Eur J Biomed Inform*, 2(1), 6-33.
- Shen, M., Zhang, J., Zhu, L., Xu, K., & Tang, X. (2019). Secure SVM training over vertically-partitioned datasets using consortium blockchain for vehicular social networks. IEEE Transactions on Vehicular Technology, 69(6), 5773-5783.
- Sitar-Taut, V. A., Zdrenghea, D., Pop, D. and Sitar-Taut, D. A. (2009), 'Using machine learning algorithms in cardiovascular disease risk evaluation', *J App Comp Sci Math*, 5(3), 29–32.
- Swain, S. K., Behera, I. C. and Sahu, M. C. (2019), 'Primary laryngeal tuberculosis: Our experiences at a tertiary care teaching hospital in Eastern India', *J Voice*, 33(5), 812-e9.
- Tangüis, H. G., Caylà, J. A., García, P., Jansà, J. M. and Brugal, M. T. (2000), 'Factors Predicting Non-Completion of Tuberculosis Treatment among HIV-Infected Patients in Barcelona (1987-1996)', *Int J Tuberc Lung Dis,* 4(1), 55- 60.
- Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O. and Poorolajal, J. (2019), 'Prediction of survival and metastasis in breast cancer patients using machine learning classifiers', *Clin Epidemiol Glob Health*, 7(3), 293-299.
- Thiam, S., LeFevre, A. M. and Hane, F (2007), 'Effectiveness of a Strategy to Improve Adherence to tuberculosis Treatment in a Resource-Poor Setting: A Cluster Randomized Controlled Trial', *JAMA*, 297(4), 380-386. https://doi.org/10.1001/jama.297.4.380.
- Vittinghoff, E, Glidden DV, Shiboski SC, McCulloch CE (2011) Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer Science & Business Media.
- Vivar, D. E. P., Cruz, Y. J. T. and Villasana, J. E. M. (2016), 'Diagnosis of extra-pulmonary tuberculosis: Systematic analysis of literature and study of seven cases in the cervicofacial region', *Revista Odontológica Mexicana,* 20(4), e258–e264.
- World Health Organization (2006), The Stop TB Strategy: Building on and enhancing DOTS to meet the TB-related Millennium Development Goals, Geneva.
- Yew, W. W. (1999), 'Directly Observed Therapy Short-Course: The Best Way to Prevent Multidrug-Resistant Tuberculosis', *Chemother*, 45(2), 26-33. https://doi.org/10.1159/000048479.
- Zhang, S. (2020). Cost-sensitive KNN classification. Neurocomputing, 391, 234-242.