How Machine Learning Can Help the Classification of Treatment Outcomes of Tuberculosis: A Systematic Review

Maicon Herverton Lino Ferreira da Silva Barros[1], Sebastião Rogerio da Silva Neto[1], Maria Gabriela de Almeida Rodrigues[2], Vanderson de Souza Sampaio[3], Patricia Takako Endo[1]
[1]Universidade de Pernambuco (UPE), {mhlfsb, srsn}@ecomp.poli.br, patricia.endo@upe.br
[2] Universidade do Estado do Amazonas (UEA), rodriguesgabriela016@gmail.com
[3] Insituto Todos pela Saúde (ITpS), vandersons@gmail.com

Abstract

Tuberculosis (TB) is a disease with a global impact that over the years has mainly affected the poorest countries. After confirming the TB diagnosis, the health professional needs to analyze the severity of the clinical situation of the patient in order to make decisions about their treatment, which may include admission to Intensive Care Unit (ICU). The aim of this paper is to present a systematic review focused on Machine Learning (ML) models for predicting TB treatment outcomes. From 253 articles found through a boolean search, only 12 of them were classified as relevant, presented and discussed in this work. Results show that the current literature is focused on binary classification, mainly using tree-based ML algorithms. Based on the results of this systematic review, we state that there are many opportunities to develop new scientific projects in this area, highlighting the need for rigorous methodology to conduct models' configuration as well as experiments to evaluate them.

Keywords: Tuberculosis, Treatment Outcomes, Machine Learning, Prediction, Prognosis.

1. Introduction

Tuberculosis (TB) is an airborne infectious disease caused by the bacillus Mycobacterium tuberculosis (Mtb) that typically affects the lungs but can also affect other parts of the body, such as the brain. According to the last World Health Organization (WHO) Global TB Report "*until the COVID-19 pandemic, TB was the leading cause of death by an infectious agent*" (WHO, 2021). Despite having a global effort conducted by the WHO to reduce the incidence of TB and its mortality rate, unfortunately, the COVID-19 pandemic has reversed years of evolution towards reducing TB infection worldwide.

Since TB is a difficult disease to eliminate, health programs around the world have focused their goal on early TB diagnosis (Martins and de Miranda, 2020). However, after confirming the TB diagnosis, it is necessary to understand the severity of the clinical situation of the patient in order to make decisions about their treatment, which may include an admission to Intensive Care Unit (ICU) or a long course drug treatment.

The complexity of predicting what may happen to a patient at the end of a TB treatment is high because it presents specific outcomes (Organization, 2013), such as *(i)* the patient having a complete treatment but the disease was not cured either for not having done a specific exam or the exam was negative (a drug-resistant case); *(ii)* the patient may have treatment failure (a non-adherence case); *(iii)* the patient may die from the disease, among others. Due this complexity, some articles in the literature have proposed the application of Machine Learning (ML) models to help healthcare professionals when making decisions related to TB treatment, in order to increase the probability of its success and, as consequence, improving the patient's quality of life. We also emphasize that TB remains a major cause of morbidity and mortality in many low- and middle-income countries, including South Africa, Nigeria, and India (Pai et al., 2016; WHO, 2021). Therefore, ML models may be presented as low-cost solutions for these cases.

However, the construction of a ML models requires a rigorous scientific process in order to produce a model capable of generalizing the data and providing satisfactory results. In this work, we present a systematic review on existing researches that employ

URI: https://hdl.handle.net/10125/102803
978-0-9981331-6-4

ML techniques to classify TB treatment outcomes.

2. Methodology

A systematic review that that adheres to the definitions in Kitchenham and Charters (2007) goes through three phases: *(i)* planning, *(ii)* conducting and *(iii)* displaying the results.

To achieve this goal, this systematic review follows the methodology presented in Figure 1 and seeks to address the following research questions:

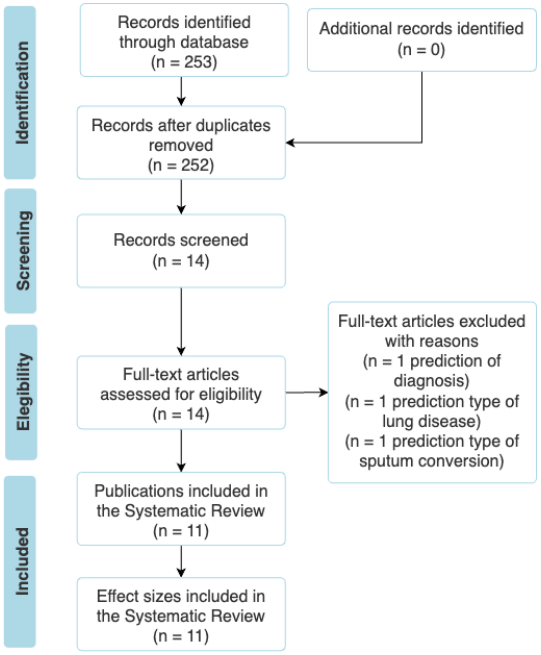

Figure 1. Methodology for article selection.

- **RQ 01**: What ML techniques are being used to classify treatment outcomes of TB?
- **RQ 02**: What are the main characteristics of the data sets used by the articles?
- **RQ 03**: What are the metrics being used to evaluate the performance of ML techniques?

2.1. Sources

We chose the following five databases, which present relevant articles in the review scope field: IEEE Xplore[1]; Google Scholar[2]; ACM Digital Library[3]; PubMed[4]; Springer[5].

[1] ieee.org
[2] scholar.google.com
[3] dl.acm.org
[4] pubmed.ncbi.nlm.nih.gov
[5] link.springer.com

2.2. Filters

This systematic review was conducted with automatic search in those five selected research sources making use of: *(i)* keywords; and *(ii)* period in years. Keywords and synonyms related to the research were combined to compose a relevant search string, in order to select articles strongly related to the topic. It was defined as follows:

["artificial intelligence" OR "deep learning" OR "machine learning" OR "neural network"] AND ["classification" OR "prediction"] AND ["prognosis" OR "treatment outcomes"] AND ["tuberculosis" OR "TB"] IN (Metadata) OR (Title) OR (Abstract), between 2012-2021.

2.3. Inclusion and exclusion criteria

The inclusion criteria were: *(i)* addressed prediction or classification problem for treatment outcomes of TB; or *(ii)* focused on classification for treatment outcomes of TB. The exclusion criteria were: *(i)* was not written in English language; *(ii)* was published as poster, tutorial or editorial; *(iii)* duplicated; and *(iv)* did not apply AI.

The boolean string was run on the search sources and initially returned 253 records. These articles had their titles and abstracts evaluated by two authors according to inclusion and exclusion criteria. When a decision conflict arose, a third author made the decision. These authors are among the authors of the manuscript and they made their decisions about inclusion or exclusion independently, without cross-referencing information between them.

2.4. Data extraction and coding

The following data were extracted for each article: authors, year of publication, forms of TB considered, age of patients, articles that use the WHO definition for predicted class, type of classification (regression, binary or multi-class), ML technique(s) employed, data set used in the work, data set balancing technique, data attributes used as input, hyperparameter optimization techniques, attribute selection techniques used, benchmark of models, metrics used to evaluate the ML performance and statistical methods used to analyze the results.

3. Results

After applying the inclusion and exclusion criteria, 253 reading articles were identified, after removing the duplicates (n=1) 252 remained. At the end, 14 articles were selected and after reading, two of them were

Table 1. Type of problem covered by selected articles.

Selected articles	Type of problem	Target classification (WHO)	Models used in the articles
A. Zhang et al. (2021)	Binary classification	DRUG-RESISTANT or NOT	DCNN
Rosenfeld et al. (2021)	Binary classification	CURED or DIED	LOR
Asad et al. (2020)	Binary classification	TREATMENT FAILURE or NOT	J48, ANN, KNN, SVM and RF
Chen et al. (2019)	Binary classification	DRUG-RESISTANT or NOT	WDNN, MLP, RF and LOR
Killian et al. (2019)	Binary classification	FAVORABLE OUTCOME, UNFAVORABLE OUTCOME	LIR, RF, SVM, Leap-LSTM
Sauer et al. (2018)	Binary classification	TREATMENT SUCCESS or TREATMENT FAILURE	LASSO, RF and SVM
Hussain and Junejo (2019)	Binary classification	TREATMENT COMPLETE or TREATMENT INCOMPLETE	SVM, RF and ANN
Gao and Qian (2017)	Binary classification	TREATMENT SUCCESS or NOT	CNN
Mburu et al. (2018)	Binary classification	TREATMENT SUCCESS or TREATMENT FAILED	CART
Kanesamoorthy and Dissanayake (2021)	Multi-class classification	NO SYMPTOMS, DRUG RESISTANCE or GET WORSE	SVM
Swaminathan et al. (2016)	Multi-class classification	TREATMENT COMPLETED, TREATMENT FAILED or DIED	CART, TreeNet and RF

excluded (one for being about diagnosis and another for making a prediction of lung cancer). Thus, 11 articles remained and were included for further analysis.

3.1. RQ 01: What ML techniques are being used to classify treatment outcomes of TB?

Determining treatment outcomes of TB is not a trivial task. For three decades, the health status and quality of life during the treatment of a disease has been increasingly receiving attention in the health area (Ware Jr, 1984). Monitoring treatment outcomes of TB is an important task that can help reduce the early mortality of a patient diagnosed with this disease (Jiménez-Corona et al., 2013).

Table 1 presents articles by type of problem they solved and the target classification. Regarding the type of classification, most of articles (nine of 11) presented models for binary classification in treatment outcomes of TB, considering the following target classes defined by WHO (Table 2).

Table 2. Treatment outcomes for TB patients (based on WHO definitions (Organization, 2013))

Outcome	Definition
CURED	A pulmonary TB patient with bacteriologically confirmed TB at the beginning of treatment who was smear- or culture-negative in the last month of treatment and on at least one previous occasion.
TREATMENT COMPLETED	A TB patient who completed treatment without evidence of failure BUT with no record to show that sputum smear or culture results in the last month of treatment and on at least one previous occasion were negative, either because tests were not done or because results are unavailable.
TREATMENT FAILED	A TB patient whose sputum smear or culture is positive at month 5 or later during treatment.
DIED	A TB patient who died for any reason before starting or during the course of treatment.
LOST TO FOLLOW-UP	A TB patient who did not start treatment or whose treatment was interrupted for 2 consecutive months or more.
NOT EVALUATED	A TB patient for whom no treatment outcome is assigned. This includes cases "transferred out" to another treatment unit as well as cases for whom the treatment outcome is unknown to the reporting unit.
TREATMENT SUCCESS	The case of cured and treatment completed.

Only two articles did not follow the WHO definition and used the following classes: DRUG-RESISTANT or NOT (Chen et al., 2019; A. Zhang et al., 2021).

According to Garcıa-Pedrajas and Ortiz-Boyer (2011), multi-class classification problem face two main issues: "*many algorithms work better with binary problems or are specifically designed for two-class problems*". Therefore, maybe due the additional complexity of this type of problem, we found only two articles focused to solve a multi-class classification: Kanesamoorthy and Dissanayake (2021) considered three classes that do not follow WHO definitions: NO SYMPTOMS, DRUG-RESISTANCE and GET WORSE; and Swaminathan et al. (2016) classified on TREATMENT COMPLETED, TREATMENT FAILED an DIED.

In this systematic review, most of the proposed models applied supervised learning, in which training data is fed into ML algorithms including the desired class, also called target class (Kang and Jameson, 2018). ML models used in selected articles were: Decision Trees (DT) family that includes TreeNet, J48, C4.5, CART and Random Forests (RF); Artificial Neural Netarticles (ANN) also known as Neural Netarticles (NN) or Multi-layer Perceptron (MLP); k-Nearest Neighbors (kNN); Support Vector Machine (SVM); Logistic and Linear Regression (LOR, LIR) (Mahesh, 2020), as shown in Table 1. DL models that were used in selected articles were: Convolutional Neural Networ (CNN), Deep Convolutional Neural Netarticles (DCNN), Wide Deep Neural Netarticles (WDNN) and Long Short-Term Memory (LSTM).

DT family was the most used technique to solve binary classification with a total of seven models proposed by (Asad et al., 2020; Chen et al., 2019; Hussain and Junejo, 2019; Killian et al., 2019; Mburu et al., 2018; Sauer et al., 2018), followed by SVM and CNN/DCNN/WDNN models that were proposed by three articles (Chen et al., 2019; Gao and Qian, 2017; A. Zhang et al., 2021).

In general, DT family models present good interpretability and tend to be more accepted by the medical community, which may have contributed to the choice of DT family models in most studies (Tjoa and Guan, 2020). Other ML techniques, unlike the DT family models, have some limitations in terms of opacity or lack of transparency in the results (Carvalho et al., 2019; Du et al., 2019), however in more critical domains such as health, a lack of interpretability or explainability can have harmful effects (Molnar, 2020).

There is growing interest among the academic community and industry in interpreting ML (Du et al.,

Table 3. Configuration of models proposed by selected articles.

Selected articles	ML and/or DL	Model configuration	Hyperparameter optimisation
A. Zhang et al. (2021)	DCNN	4 convolutional layers, each with 4 x4 filters to extract features from the input genomic images. The last convolutional layer uses a stride to reduce the genomic image dimension, which is then reshaped to a 1-dimensional array before a fully connected layer that generates the output	Grid Search
Rosenfeld et al. (2021)	Logistic Regression	Univariate and multivariate logistic regression	None
Chen et al. (2019)	WDNN	Not described	Bayesian optimization
	MLP		
	RF		
	Logistic regression		
	Linear Regression	Not described	Grid Search
	RF	150 trees	
	SVM	Not described	
Killian et al. (2019)	Leap-LSTM	64 hidden units for LSTM, 48 units dense layers, 4 units in penultimate dense layer	
Sauer et al. (2018)	RF	Not described	None
	SVM		
	SVM	Gaussian kernel (radial) with cost value 6.	
	RF	RF, chose the number of trees to grow up to 2000 trees, and number of variables randomly sampled at each split as 26.	
Hussain and Junejo (2019)	ANN	The number of hidden layers, weight decay parameter, and maximum iterations to stop was chosen as 3, 0.9, and 1000, respectively.	Employ stratified sampling technique to ensure that the ratio of the two classes remained the same in the three sets.
Gao and Qian (2017)	CNN	MatConvNet from Matlab was used. Six CNN layers are designed with input data of 64 × 64 pixels. The filter sizes for each layer are (4, 4), (3, 3), (3, 3), (2, 2), (2, 2) and (3, 3), respectively.	None
Mburu et al. (2018)	CART	Boosted CART modeling was performed using Tree Net in Salford software version 8	None
Kanesamoorthy and Dissanayake (2021)	SVM	Not described	None
	CART, TreeNet		
Swaminathan et al. (2016)	RF	Not described	None
Asad et al. (2020)	J48	Not described	Random choices made by the author.
	ANN	Tanh activation function proves to be useful with lbfgs solver, 1-8 hidden layer.	
	KNN	Euclidean distance is used to measure the distance between the data points for k = 7.	
	SVM	Linear and polynomial kernels.	
	RF	1 to 10 trees	

2019) results. EXplainable Artificial Intelligence (XAI) is one of the solutions to cover this limitation, as it provides interpretability or explainability of the behavior of these techniques, in a comprehensive way for humans (Carvalho et al., 2019). Its adoption is beneficial for end users because it encourages the adoption of ML based systems and gives them more trust thanks to the interpretability of these approaches (Du et al., 2019).

3.2. Hyperparameter optimization

As stated by Feurer and Hutter (2019), automated hyperparameter optimization has several important use cases: *(i)* reduce the human effort required for configuring a ML model; *(ii)* improve the performance of ML models; and *(ii)* improve the reproducibility of scientific studies. However, we emphasize that most of the articles found in this systematic review did not conduct any hyperparameter optimization, as shown in Table 3.

Only two articles performed Grid Search (Killian et al., 2019; W. Zhang et al., 2018). Grid search, also known as full factorial design, is a simple hyperparameter optimization in which a finite set of values for each hyperparameter is defined and the Cartesian products of these sets are evaluated in an exhaustive way. Despite the simplicity, Grid Search suffers from dimensionality, since the number of evaluations grows exponentially with the number of hyperparameter values defined for the experiment. Chen et al. (2019) applied Bayesian optimization. According to Yang and Shami (2020), compared with traditional optimization methods, Bayesian optimization models can be more suitable for hyperparameter optimization. It is based on a probabilistic surrogate model of the objective function that determines the next hyperparameter value based on previously-obtained results, avoiding unnecessary evaluations.

3.3. RQ 02: What are the main characteristics of the data sets used by the articles?

Table 4 summarizes the data sets used by the selected articles in this systematic review by number of records, number of original attributes, number of attributes used by the proposed model, location where data were collected and whether they applied any feature selection technique.

3.3.1. Feature selection Feature selection is one of the most important step of a classic ML model development process (Galelli et al., 2014), since attributes contain the information necessary for the

Table 4. Characteristics of the data sets used to evaluate ML models for TB outcome treatment.

Selected articles	Records	Number of attributes in data set	Number of attributes used	Country/Location of data set	Feature selection technique	Data availability
A. Zhang et al. (2021)	149	Not described	Not described	Not described	Not described	Not described
Rosenfeld et al. (2021)	253	14	5	Belarus	ANOVA and Chi-square	Not publicly available
Chen et al. (2019)	3,601	Not described	222	Not described	Not described	Publicly available
Killian et al. (2019)	4,167	Not described	29	India	Not described	Not described
Sauer et al. (2018)	643	23	23	Azerbaijan, Belarus, Georgia, Moldova, Romania	FSS, BSE, BEFSS	Not publicly available
Hussain and Junejo (2019)	4,213	84	52	Stop TB partnership	Chi-square test	Partial publicly available
Gao and Qian (2017)	230	Not described	Not described	Dublin (collected) origin not described	Not described	Not publicly available
Mburu et al. (2018)	340	Not described	10	Kenya	CART	Not publicly available
Kanesamoorthy and Dissanayake (2021)	356	35	20	Myanmar, Yangon	GA	Publicly available
Swaminathan et al. (2016)	143	30	15	India	Not described	Not described
Asad et al. (2020)	1,295	22	11	Azerbaijan, Belarus, Georgia, Moldova, Romania	Univariate Feature Selection (UFS)	Not publicly available

models to be able to somehow generalize the learning of a given problem.

At same time, feature selection techniques can also be used to reduce the computational cost of training, and in some cases to improve the performance of the model. ML models are highly sensitive to the attributes selected as input for their training (Fernando et al., 2009).

Figure 2 presents the most frequent attributes used by selected articles as input for their models. We found 157 different attributes, but for a better graphical presentation, only attributes with a frequency greater than or equal to two were displayed. Gender, age, gene and weight were the most frequent attributes. Many attributes are laboratory data and a few others are socio-demographic.

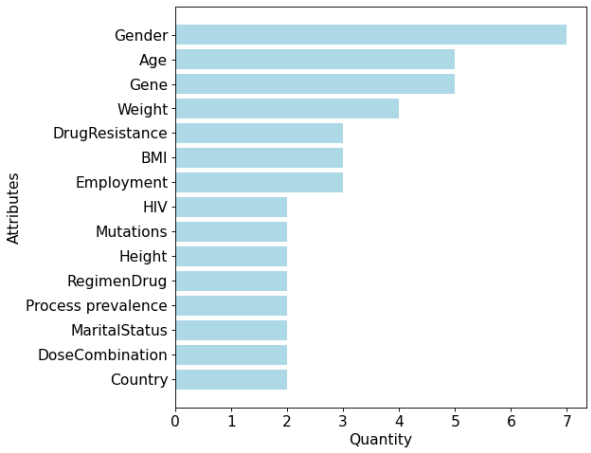

Figure 2. Frequency of attributes used by selected articles.

Only A. Zhang et al. (2021) did not describe the number of attributes used as input for their proposed models. Although Chen et al. (2019) have used 222 attributes, characterizing a model with a high dimensionality, it positively impacted the performance of their proposed models, presenting a good AUC of 98.40% and 98.20%. On the other hand, Rosenfeld et al. (2021) used far fewer attributes, only 5, and also obtained a good result in metrics BACC (82.00%) and MCC (0.39).

Swaminathan et al. (2016) used only 15 of the 30 attributes present in the original data set, but authors did not describe whether they used any feature selection technique and obtained 77.00% by CART model and 75.00% by RF model in ROC metric.

Sauer et al. (2018) applied three wrapper feature selection techniques: Forward Stepwise Selection (FSS), Backward Stepwise Elimination(BSE), Backwards Elimination and Forward Stepwise Selection (BEFSS). They used data set comprising 643 patients records from five country: Azerbaijan, Belarus, Georgia, Moldova and Romania. The data set has 23 attributes, and all of them were considered for training and testing the models. Note that authors used feature selection techniques to identify the importance of attributes, and not to reduce the dimensionality of the set of evaluated attributes.

Kanesamoorthy and Dissanayake (2021) also applied wrapper feature selection techniques, the Genetic Algorithm (GA), in a data set that was originally composed of 35 attributes and 20 of them were selected. GA mimics the biological process of reproduction and natural selection to solve mainly optimization problems (Kinnear et al., 1994; Reading and Wesley, n.d.).

Hussain and Junejo (2019) and Asad et al. (2020) used filter techniques: Chi-square test and Univariate Feature Selection (UFS), respectively. Chi-square test, according to Kothari (2013), is used to check if there is a correlation between non-numeric variables. The UFS uses univariate analysis to extract attributes from

Table 5. Distribution of samples per classes.

Selected articles	Target classification	Samples	Proportion
A. Zhang et al. (2021)	DRUG-RESISTANT - RESISTANT STRAINS	75	0,503
	DRUG-RESISTANT - SUSCEPTIBLE STRAINS	74	0,497
Rosenfeld et al. (2021)	CURED	228	0,902
	DIED	25	0,098
Kanesamoorthy and Dissanayake (2021)	NO SYMPTOMS, DRUG RESISTANCE and GET WORSE	356	1.000
Asad et al. (2020)	TREATMENT FAILURE or NOT	1,295	1.000
Chen et al. (2019)	DRUG-RESISTANT or NOT	3,601	1.000
Killian et al. (2019)	FAVORABLE OUTCOME	3,734	0,896
	UNFAVORABLE OUTCOME	433	0,104
Sauer et al. (2018)	TREATMENT SUCCESS	491	0,764
	TREATMENT FAILURE	152	0,236
Hussain and Junejo (2019)	TREATMENT COMPLETE	2,712	0,644
	TREATMENT INCOMPLETE	1,501	0,356
Gao and Qian (2017)	TREATMENT SUCCESS (DRUG SENSITIVE)	134	0,583
	TREATMENT SUCCESS (MULTIDRUG-RESISTANT)	96	0,417
Mburu et al. (2018)	TREATMENT SUCCESS	308	0,906
	TREATMENT FAILED	32	0,094
Swaminathan et al. (2016)	TREATMENT COMPLETED	110	0,769
	TREATMENT FAILED	24	0,168
	DIED	9	0,063
	11	0,393	

*Numbers were rounded.

the data set, and it is often used with data sets that have many attributes (Asad et al., 2020; D'Agostino, 2017).

As can be seen in Table 4, six articles did not described if they applied feature selection mechanism (Chen et al. (2019), Gao and Qian (2017), Mburu et al. (2018), Rosenfeld et al. (2021), Swaminathan et al. (2016) and A. Zhang et al. (2021)). Even Rosenfeld et al. (2021) and Swaminathan et al. (2016) that explicitly made some feature selection did not described how those features were sort out, demonstrating the difficult of reproducing their experiments.

Principal component analysis (PCA) is a very common technique usually applied for feature selection (Kurita, 2019), but it was not used in any of the articles found in this systematic review.

Only three articles provided information about how to access their data set: Chen et al. (2019)[6], Hussain and Junejo (2019)[7], Kanesamoorthy and Dissanayake (2021)[8], and two articles made their source code publicly available: Hussain and Junejo (2019) and Rosenfeld et al. (2021). Having access to the data set and/or the code of the experiment is essential for the research to be replicated and for a more technical analysis of model performance.

[6]www.reseqtb.org. This link was not available when writing this manuscript.

[7]github.com/seekme94/tuberculosis-predictive-analytics/blob/master/sample_dataset.csv

[8]journals.plos.org/plosone/article/file?type=supplementary&id=10.1371/journal.pone.0177999.s001

3.3.2. Data set size Only four articles presented a data set having more than 1,000 records: Chen et al. (2019), Killian et al. (2019), Asad et al. (2020) and Hussain and Junejo (2019). Hussain and Junejo (2019) used the biggest data set, comprising 4,213 patients records from Stop TB, that is a global partnership to control TB that implements mechanisms to TB cured and sets goals to ensure quality and effective treatment.

A negative aspect of some articles is the absence of information about the data set used. Here, we highlight two articles: A. Zhang et al. (2021) and Gao and Qian (2017) that did not mention the original attributes of the data set, neither the set of attributes used to train their models. Chen et al. (2019), Killian et al. (2019) and Mburu et al. (2018) mentioned the final features but did not present the original data set. As said previously, the lack of a clear methodology imposes very hard to replicate, validate and compare their articles.

Table 5 shows the percentage relationship between how balanced or imbalanced the data set is, regarding the target classes. Three articles did not described the amount of records per target class: Kanesamoorthy and Dissanayake (2021); Asad et al. (2020) and Chen et al. (2019). With the exception of the article presented by A. Zhang et al. (2021), which has similar proportion of records for each target class, none of other articles explicitly describe how (or if) they deal with imbalanced data.

The most imbalanced data set was used by Rosenfeld et al. (2021) with 228 records of the CURED class,

that represent 90.20% of the total records in the data set. Swaminathan et al. (2016) used a data set with multi-classes, where 6,30% of the records belong to the DIED class; 76,90% of TREATMENT COMPLETED and 16,80% of TREATMENT FAILED.

According to Chicco and Jurman (2020), evaluation metrics influence the interpretation of models' performance if data sets are imbalanced. It means that if the aforementioned articles did not use adequate metrics, there may be biases in the results and discussions presented in their articles (see Section 3.4).

3.4. RQ 03: What are the metrics being used to evaluate the performance of ML techniques?

Figure 3 presents the metrics used to evaluate the proposed models in the literature. Some articles used more than one metric and are duplicated in the graph. Sensitivity and accuracy were the most used metrics.

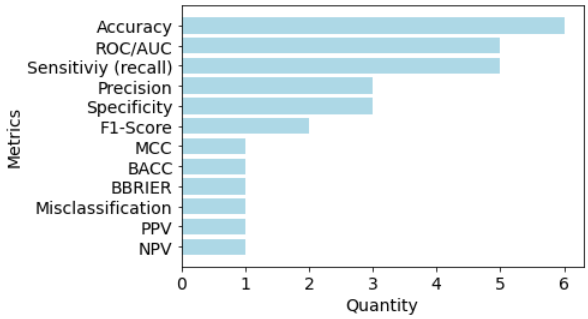

Figure 3. Metrics used to evaluate the models proposed in the literature.

We highlight the Matthews Correlation Coefficient (MCC) metric, that was used by Rosenfeld et al. (2021). This metric ranges [-1, 1] and a result close to 1 indicates a very good prediction. There is a strong positive correlation of the prediction and TP. This high correlation indicates that the variables strongly agree. When MCC is equal to 0, there is no correlation between variables, the classifier randomly assigns units to classes with no real value (Boughorbel et al., 2017; Grandini et al., 2020). The usage of MCC metric is the most indicated when one wants to obtain a metric that evaluates the model in general (the four categories of the confusion matrix) regardless of imbalancing of data (Chicco and Jurman, 2020).

3.5. Models' performance

Table 6 presents the performance results of all models proposed by the selected articles. We have presented only the best result, by metric and model, among the experiments performed by articles considered in this systematic review. The best accuracy, 93.10%, was achieved by the DCNN model proposed by A. Zhang et al. (2021), that was created using genomic data of bacterial strains converted into genomic images with the objective to predict the resistance or not to the drug pyrazinamide.

Note that Rosenfeld et al. (2021) obtained good results in four metrics (BACC, BBRIER, MCC and sensitivity). However, their Logistic regression model obtained 100% of sensitivity, and at same time, 0.00% of specificity, which probably represents a bias in the model learning. This issue was reported by the authors: "*while no class balancing had higher overall accuracy and sensitivity (optimization led to prediction similar to a featureless model where only cured outcome is predicted*".

As sensitivity and specificity are inversely proportional metrics, in binary classification, the discrepancy between them is an indication that the model is classifying all examples of one of the class (for example, CURED) and is missing all instances of the other class (E.g DIED). This also occurs when the data set is imbalanced and the model cannot generalize both classes. One of the solutions is to balance the data set or use appropriate metrics for imbalanced data sets, as authors also did experiments like these.

Overall, binary-classified models performed better compared to multi-class models. Taking into account the DT family models, they obtained a good accuracy, but the DCNN model proposed by A. Zhang et al. (2021) outperformed them. It is also important to note that Chen et al. (2019) obtained good results with their LOR model (98.40% AUC and 96.10% sensitivity) and with their WDNN model, even having used a high dimensionality (222 attributes). However, they did not explain details of their model settings.

4. Discussion

This systematic review presented the current state-of-the-art on usage of ML models to support healthcare professionals in treatment outcomes of TB. Comparing articles is not a trivial task, due to the variation in their focus, such as the type of classification or regression problem, the target class of classification/prediction and configurations of ML models.

We would like to report an issue related to scientific reproducibility. As presented in Table 3, four articles (Chen et al., 2019; Kanesamoorthy and Dissanayake, 2021; Sauer et al., 2018; Swaminathan et al., 2016) did

Table 6. Results from evaluate metrics used by selected articles.

		Best results by metrics and by model from selected articles									
Selected articles	Model	Accuracy	AUC	ROC	BACC	BBRIER	MCC	Precision	F1-score	Sensitivity	Specificity
A. Zhang *et al.* (2021)	DCNN	93.10%	–	–	–	–	–	–	–	–	–
Rosenfeld *et al.* (2021)	LOR	92.00%	84.00%	–	82.00%	91.00%	0.39	–	–	100.00%	67.00%
Asad *et al.* (2020)	J48	92.00%	–	–	–	–	–	–	–	–	–
	ANN	89.00%	–	–	–	–	–	–	–	–	–
	KNN	90.00%	–	–	–	–	–	–	–	–	–
	SVM	91.00%	–	–	–	–	–	–	–	–	–
	RF	92.00%	–	–	–	–	–	–	–	–	–
	WDNN	–	98.20%	–	–	–	–	–	–	95.40%	97.80%
	MLP	–	–	–	–	–	–	–	–	–	–
	RF	–	–	–	–	–	–	–	–	–	–
Chen *et al.* (2019)	LOR	–	98.40%	–	–	–	–	–	–	96.10%	100%
Killian *et al.* (2019)	LIR	–	–	–	–	–	–	–	–	–	–
	RF	–	80.50%	–	–	–	–	–	–	–	–
	SVM	–	–	–	–	–	–	–	–	–	–
	Leap-LSTM	–	79.30%	–	–	–	–	–	–	–	–
	LASSO	–	72.00%	–	–	–	–	–	–	21.00%	96.00%
	RF	–	70.00%	–	–	–	–	–	–	30.00%	91.00%
Sauer *et al.* (2018)	SVM	–	69.00%	–	–	–	–	–	–	21.00%	94.00%
Hussain and Junejo (2018)	SVM	74.23%	–	–	–	–	–	73.05%	–	44.04%	95.71%
	RF	76.32%	–	–	–	–	–	67.55%	–	62.31%	86.07%
	ANN	76.01%	–	–	–	–	–	49.13%	–	68.50%	78.53%
Gao and Qian (2018)	CNN	91.11%	–	–	–	–	–	–	–	–	–
Mburu *et al.* (2018)	CART	–	–	65.00%	–	–	–	–	–	–	–
Kanesamoorthy and Dissanayake (2021)	SVM	67.00%	–	–	–	–	–	67.00%	57.00%	–	–
Swaminathan *et al.* (2016)	CART	–	–	77.00%	–	–	–	–	–	–	–
	TreeNet	–	–	–	–	–	–	–	–	–	–
	RF	–	–	75.00%	–	–	–	–	–	–	–

not provide information about their proposed models, and in Table 4 some other articles (Gao and Qian, 2017; A. Zhang et al., 2021) did not describe the attributes used to train their models. In these cases, it is very hard (and sometimes unfair) discuss and compare the results of their articles because there is no subsidy to clarify and justify the performance of the proposed models. The lack of such information shows the weakness of the scientific method applied in the articles. Therefore, having these data and source code available may improve the quality of future articles, turning the research more accessible and reproducible.

A article to be highlighted is Rosenfeld et al. (2021) that classifies treatment outcomes (prognosis) of TB between CURED and DIED classes. It is important to note that the data set used is clearly imbalanced with very few CT scan records with only 25 patients who TB DIED, and 228 patients TB CURED. However, it is also worth mentioning that authors used adequate metrics that are not sensitive to the imbalance of classes such as Balanced Accuracy (BACC) and MCC. According to Chicco and Jurman (2020), MCC is a metric that is not affected by the problem of imbalanced data sets and is the only binary ranking metric that generates a high score only if the binary predictor was able to correctly predict most positive data instances and most negative data instances. MCC result ranges from -1 and 1. BACC can be defined as the arithmetic mean of sensitivity and specificity. It weighs the two classes equally, taking into account the positive and negative errors caused by the imbalanced of the classes. In addition, the authors benchmark the models used, including applying class balancing techniques. Rosenfeld et al. (2021) obtained 0.39 MCC with balanced class. However, it is worth remembering that 25% of the data were used for these tests, that is, only 6 records of TB DIED. Although there is no apparent bias in the models, there is no way to guarantee that the proposed model can generalize, as the number of training examples were low as well as the examples used for testing, which may be insufficient for generalizing a problem as complex as classifying treatment outcomes of TB. One way to get around this problem is to increase the number of records (examples) for the classes, in order to compare different scenarios with different numbers of records. There are some alternatives to increase minority classes such as the Synthetic Minority Oversampling Technique (SMOTE), which carries out an interpolation among neighboring minority class instances, helping classifiers to improve the generalization ability (Fernández et al., 2018). In this way, it is possible to evaluate and discuss the generalization capacity of an ML model.

Based on results of this systematic review, future articles can explore the effectiveness of different approaches for classifying treatment outcomes of TB, with real data in order to evaluate the generalization of the model.

Conclusions

This systematic review presented articles from literature that apply ML to cover the classification of treatment outcomes of TB. Based on the selected articles, we noted that the usage of ML is still premature, presenting many possibilities to develop new research projects in this area.

Most articles aimed to classify possible failure during the TB treatment or drug resistance, and the most common evaluation metrics were accuracy, AUC, ROC, precision, sensitivity (recall) and specificity, despite evidence showing that the data sets used by most articles are imbalanced, what may causes bias when using inappropriate metrics to test the model.

The usage of ML models to classify treatment outcomes of TB is still very preliminary and with few data used to date, showing that there are many research opportunities in this area.

Acknowledgements

This work was partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE), and Universidade de Pernambuco (UPE), an entity of the Government of the State of Pernambuco focused on the promotion of teaching, research and extension.

References

Asad, M., Mahmood, A., & Usman, M. (2020). A machine learning-based framework for predicting treatment failure in tuberculosis: A case study of six countries. *Tuberculosis*, *123*, 101944.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, *12*(6), e0177678.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chen, M. L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., Kohane, I. S., Beam, A., & Farhat, M. (2019). Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in mycobacterium tuberculosis resistance prediction. *EBioMedicine*, *43*, 356–369.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13.

D'Agostino, R. (2017). *Goodness-of-fit-techniques*. Routledge.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

Fernando, T., Maier, H., & Dandy, G. (2009). Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology*, *367*(3-4), 165–176.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning* (pp. 3–33). Springer, Cham.

Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., & Gibbs, M. S. (2014). An evaluation framework for input variable selection algorithms for environmental data-driven models. *Environmental Modelling & Software*, *62*, 33–51.

Gao, X. W., & Qian, Y. (2017). Prediction of multidrug-resistant tb from ct pulmonary images based on deep learning techniques. *Molecular pharmaceutics*, *15*(10), 4326–4335.

Garcıa-Pedrajas, N., & Ortiz-Boyer, D. (2011). An empirical study of binary classifier fusion methods for multiclass classification. *Information Fusion*, *12*(2), 111–130.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*.

Hussain, O. A., & Junejo, K. N. (2019). Predicting treatment outcome of drug-susceptible tuberculosis patients using machine-learning models. *Informatics for Health and Social Care*, *44*(2), 135–151.

Jiménez-Corona, M. E., Cruz-Hervert, L. P., Garcıa-Garcıa, L., Ferreyra-Reyes, L., Delgado-Sánchez, G., Bobadilla-del-Valle, M., Canizales-Quintero, S., Ferreira-Guerrero, E., Báez-Saldaña, R., Téllez-Vázquez, N., et al. (2013). Association of diabetes and tuberculosis: Impact on treatment and post-treatment outcomes. *Thorax*, *68*(3), 214–220.

Kanesamoorthy, K., & Dissanayake, M. B. (2021). Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. *International Journal of Mycobacteriology*, *10*(3), 279.

Kang, M., & Jameson, N. J. (2018). Machine learning: Fundamentals. *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, 85–109.

Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., & Tambe, M. (2019). Learning to prescribe interventions for tuberculosis patients using digital adherence data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2430–2438.

Kinnear, K. E., Langdon, W. B., Spector, L., Angeline, P. J., & O'Reilly, U.-M. (1994). *Advances in genetic programming* (Vol. 3). MIT press.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kothari, C. (2013). *Quantitative techniques (new format)*. Vikas Publishing House.

Kurita, T. (2019). Principal component analysis (pca). *Computer Vision: A Reference Guide*, 1–4.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381–386.

Martins, V. d. O., & de Miranda, C. V. (2020). Diagnóstico e tratamento medicamentoso em casos de tuberculose pulmonar: Revisão de literatura. *Revista Saúde Multidisciplinar*, *7*(1).

Mburu, J. W., Kingwara, L., Ester, M., & Andrew, N. (2018). Use of classification and regression tree (cart), to identify hemoglobin a1c (hba1c) cut-off thresholds predictive of poor tuberculosis treatment outcomes and associated risk factors. *Journal of clinical tuberculosis and other mycobacterial diseases*, *11*, 10–16.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Organization, W. H. (2013). *Definitions and reporting framework for tuberculosis – 2013 revision: Updated december 2014 and january 2020*.

Pai, M., Behr, M., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C., & Raviglione, M. (2016). Tuberculosis. nature reviews disease primers, 2, 16076.

Reading, M., & Wesley, A. (n.d.). Freund, je (1962). mathematical statistics. englewood cliffs, nj.: Prentice-hall. goldberg, d., & richardson, j.(1987). genetic algorithms with sharing for multimodal function optimization. in proceedings of the second international conference on genetic algorithms, 148–154, san mateo, ca. morgan kaufmann. goldberg, de (1989). genetic algorithms in search, optimization, and machine learning.

Rosenfeld, G., Gabrielian, A., Wang, Q., Gu, J., Hurt, D. E., Long, A., & Rosenthal, A. (2021). Radiologist observations of computed tomography (ct) images predict treatment outcome in tb portals, a real-world database of tuberculosis (tb) cases. *Plos one*, *16*(3), e0247906.

Sauer, C. M., Sasson, D., Paik, K. E., McCague, N., Celi, L. A., Sanchez Fernandez, I., & Illigens, B. M. (2018). Feature selection and prediction of treatment failure in tuberculosis. *PloS one*, *13*(11), e0207491.

Swaminathan, S., Pasipanodya, J. G., Ramachandran, G., Hemanth Kumar, A., Srivastava, S., Deshpande, D., Nuermberger, E., & Gumbo, T. (2016). Drug concentration thresholds predictive of therapy failure and death in children with tuberculosis: Bread crumb trails in random forests. *Clinical Infectious Diseases*, *63*(suppl_3), S63–S74.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793–4813.

Ware Jr, J. E. (1984). Conceptualizing disease impact and treatment outcomes. *Cancer*, *53*, 2316–2323.

WHO. (2021). Global tuberculosis report 2021 [Accessed: 2021-01-25].

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.

Zhang, A., Teng, L., & Alterovitz, G. (2021). An explainable machine learning platform for pyrazinamide resistance prediction and genetic feature identification of mycobacterium tuberculosis. *Journal of the American Medical Informatics Association*, *28*(3), 533–540.

Zhang, W., Yang, G., Lin, Y., Ji, C., & Gupta, M. M. (2018). On definition of deep learning. *2018 World automation congress (WAC)*, 1–5.