 doi:10.20944/preprints202103.0284.v1

Article

Benchmarking of machine learning models to assist the prognosis of tuberculosis

Maicon Herverton Lino Ferreira da Silva Barros[1], Geovanne Oliveira Alves[1], Lubnnia Morais Florêncio Souza[1], Elisson da Silva Rocha[1], João Fausto Lorenzato de Oliveira[1], Theo Lynn[2], Vanderson Sampaio[3] and Patricia Takako Endo[1]*

[1] Universidade de Pernambuco, Pernambuco, Brazil; {mhlfsb, goa, lmfs, esr2}@ecomp.poli.br, {fausto.lorenzato, patricia.endo}@upe.br
[2] Dublin City University, Dublin, Ireland; theo.lynn@dcu.ie
[3] Fundação de Medicina Tropical, Amazonas, Brazil; vandersons@gmail.com
* Correspondence: patricia.endo@upe.br

Abstract: Tuberculosis (TB) is an airborne infectious disease caused by organisms in the *Mycobacterium tuberculosis (Mtb)* complex. In many low and middle-income countries, TB remains a major cause of morbidity and mortality. This work performs a benchmarking of machine learning models using a Brazilian health database related to TB confirmed cases and deaths, named SINAN-TB. The goal is to predict the probability of death by TB, assisting the TB prognosis and decision taking process. The database originally has 130 features, and many of these features had missing data, or incorrect data regarding the notification dates or birth dates, or were not related to the clinical and laboratory data. These data are treated, and after the preprocessing step, a new database with 38 features and 24,015 records is generated, having 22,876 TB cases and 1,139 deaths by TB. We design two experiments to investigated how the data unbalancing impacts on the models performance. With the evaluation of the f1-macro metric, we verify that the best result is achieved when using the imbalanced database, with the ensemble model that is composed of gradient boosting (GB), random forest (RF) and multi-layer perceptron (MLP) models.

Keywords: machine learning, benchmarking, tuberculosis, prognosis

1. Introduction

Tuberculosis (TB) is an airborne infectious disease caused by organisms in the Mycobacterium tuberculosis complex (Mtb). In many low-income and middle-income countries, such as, South Africa, Nigeria and India a TB continues to be a major cause of morbidity and mortality [1][2]. Despite the World Health Organization (WHO) global efforts to reduce the incidence of TB and its mortality rate, 10 million people fell ill with TB and 1.2 million deaths were registered in 2019 [2]. In the same year, Brazil registered 96 thousand cases of TB with a mortality rate of 7% [3]. In addition, Brazil is one of countries with the highest TB rates in the world [4].

The Brazilian *Sistema Único de Saúde* (SUS) is one of the largest public health systems in the world; being responsible for providing primary care services, from medium to high complexity, such as blood donation, chemotherapy, organ transplantation, among others [5]. SUS also provides free vaccines and medicines for people with diabetes, arterial hypertension, HIV, Alzheimer's and others. In addition, SUS attends emergencies through the *Serviço de Atendimento Móvel de Urgência* (SAMU) [6]. Currently, more than three-fourths of Brazil population rely exclusively on SUS health services [7].

The *Sistema de Informação de Agravos de Notificação* (SINAN) is a SUS system mainly composed of notifications of diseases cases present in the national list of compulsory notification. These data are routinely generated by the Epidemiological Surveillance System. SINAN has a database with clinical and laboratory data on TB patients, the SINAN-TB, that could be used in order to propose solutions in the TB prognosis.

The search for early diagnosis of the disease is a goal of health programs around the world, because it is a difficult disease to eliminate [8]. To assist this process, the literature

Citation: Lastname, F.; Lastname, F.; Lastname, F. Benchmarking of machine learning models to assist the prognosis of tuberculosis. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

 doi:10.20944/preprints202103.0284.v1

proposes the use of machine learning and deep learning models for the diagnosis of TB [9]. Literature work proposes the use of deep learning for the diagnosis of TB as in [10], [11] and [12]. Other studies are based on microscopic images for the use of convolutional neural networks (CNN) in the works of the authors [13], [14] and [13]. There are also studies related to effective treatment of TB with the elderly and people with comorbidities in [15], [16] and [17].

However, there is a lack of studies on tuberculosis prognosis. After the diagnosis confirmation, it is necessary to understand the severity of the clinical situation in order to make decisions about the treatment, including hospitalisation or admission to an Intensive Care Unit (ICU). This analysis of severity is essential to avoid worsening the disease, improve the quality of life of patients and management of health resources. Recent studies have proposed DL models to predict mortality by TB, but in other contexts, such as in the elderly population [15], morbidity due to food contamination and gastrointestinal infections [16], and postoperative in-hospital mortality [17].

The main focus of our work is the TB prognostic by predicting the probability of death. For that, we evaluate machine learning techniques in order to find the best candidate to assist health professionals. The following machine learning models are evaluated: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF) and Multilayer Perceptron (MLP). For that, the benchmarking is based on a rigorous methodology that (a) identifies the most relevant features using feature selection techniques; (b) applies a randomized search technique to select the parameters of the machine learning models that achieve the best performance results; (c) perform a benchmarking of the machine learning models; and (d) creates an ensemble model [18] [19] [20] that combines two or more machine learning models in order to find better results and reduce the risk of employing an inappropriate model.

2. Background

2.1. Feature selection techniques

The feature selection techniques are algorithms to select a subset of features from the original feature set [21]. The Sequential Forward Selection (SFS) is a greedy search algorithm that selects the feature set following a bottom-up search procedure. The algorithm starts from an empty set and this set is filled iteratively [22]. It is widely used because it is simple and fast [23]. In this work, at each iteration, a new feature is added to the feature set and the selection is based on the f1-score, which can be used to evaluate the accuracy of predictions in two-classes classification problems [24]. The stop criteria is 17 features as per [9].

2.2. Machine learning models

ML can be understood as the union of forces between statistics and computer science [25]. According to [25], ML is the basis of artificial intelligence. It is a learning process in which, using past data, it is possible to create a mathematical model to predict results or define a classification. These models can be used in the health field to identify causes, risk factors, effective treatments for diseases, among others applications [26].

In this work, we use the following machine learning techniques: LR, LDA, KNN, DT, GB, RF, MLP and ensemble models, described in the next subsections.

2.2.1. Logistic regression (LR)

LR is a machine learning technique used to build classification models [27]. In LR, it is possible to test whether two variables are linearly related and calculate the strength of the linear relationship [28]. It provides a simple and powerful method for solving a wide range of problems. For instance, in health area, logistic regression can be used to predict the risk of developing a particular disease based on the observed feature of the patient [27].

 doi:10.20944/preprints202103.0284.v1

2.2.2. Linear discriminant analysis (LDA)

LDA is a data analysis method proposed by R. Fisher. The technique works with a smaller subset of data, comparing it with the size of the original data sample, in which the data of the original problem is separable [29]. The LDA is able to deal with the problem of imbalance between the classes of the data set, and maximizes the proportion of the variance, between classes, for the variance within the class in any data set, thus ensuring maximum separability [30].

2.2.3. k-nearest neighbors (KNN)

The KNN can be used for classification and regression. The k in KNN refers to the number of nearest neighbors the classifier will retrieve and use to make its prediction [31]. It is a non-parametric classification method. In order for a d data record to be classified, its k closest neighbors are taken into account, and this forms a neighborhood of the d data [32].

2.2.4. Naive Bayes (NB)

A NB classifier is a probabilistic model based on the Bayes rule [33] along with an independence assumption, was developed in 1960 [34]. NB assigns the most likely class to a given example described by its characteristic vector. The learning of these classifiers assumes that the features are independent of a certain class [35].

2.2.5. Decision trees (DT)

DT are used to solve both classification and regression problems in the form of trees that can be incrementally updated by splitting the dataset into smaller datasets [31]. For each new element in the test set, the decision tree must be traversed from the root to one of its leaves, thus, each node in the tree must be checked, and depending on the value, it must be assigned to one of the sub-trees until that the element reaches a leaf node [36].

2.2.6. Support vector machine (SVM)

The SVM is a set of supervised learning methods that analyse data and recognize patterns, and it is used for the classification and regression analyses [37]. SVM is based on the structural risk minimization criterion and its goal is to find the optimal separating hyperplane where the separating margin should be maximized. This approach improves the generalization ability of the learning machine and solves some problems like non-linear, high dimension data separation and the classification issue that lacking of prior knowledge [38].

2.2.7. Gradient boosting (GB)

GB is an iterative ensemble procedure for supervised tasks (classification or regression) which combines multiple weak-learners to create a strong ensemble [39]. In GB the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble [40].

2.2.8. Random forest (RF)

RF is currently one of the most used machine learning algorithms among mining techniques, as it is a technique that can be used for both prediction and classification and are easy for training. This preference is attributable to its high learning performance and low demands with respect to input preparation and hyper-parameter tuning [41]. Basically it is a technique that unifies several decision trees referring to the input data of the database. Thus, the classifier consists of N trees, where N is the number of trees to be cultivated, which can be any user-defined value. To classify a new data set, each case of the data sets is passed to each of the N trees. The forest chooses a class with the maximum N votes [1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1

2.2.9. Multilayer perceptron (MLP)

The MLP is machine learning model in use both for classification and regression [42]. Basically it is a perceptron model with one or more hidden layers, each layer having a certain amount of neurons, which are connected by weights. The data of the independent variables are inserted in the neurons of the input layer and are processed in the hidden layer, finally, in the output layer we have the result of the MLP.

2.3. Ensemble

Ensemble methods train several machine learning models to solve the same problem. In contrast to a single classifier, ensemble methods try to build a set of models and combine them. Ensemble learning is also called learning based on committees or multiple learning classifier systems [43].

The combination of the learning models, can be made traditionally in three ways: by average, by vote or by learning model. By average is generally applied when handling numerical outputs, and average of the values obtained as output by the classifiers is calculated. By vote, a count is made from the outputs of the classifiers based on the frequency of appearances of a class, and the class with the highest number of votes is used as an input for a new learning model [43].

2.4. Randomized search

The randomized search technique sets up a grid of hyper-parameter values and selects random combinations to train a given model [44]. Randomized search can outperform a grid search [44], especially if only a small number of hyper-parameters is used to compare the performance of the machine learning model.

2.5. Wilcoxon hypothesis test

The Wilcoxon test is a non-parametric test used to test the hypothesis that the probability distribution of the first sample is equal to the probability distribution of the second sample [45]. In this work, the Wilcoxon test is used to eliminate models that have similar distributions and then select machine learning models to compose the ensemble model.

2.6. Evaluation metric

In this study, two metrics are used, f1-score and f1-macro [46][47]. To understand these metrics, it is important to define the composition of a confusion matrix: true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Two other metrics are used to form our main assessment measures, recall and precision. Recall indicates the correct classifications among all expected cases as correct, that is, the ratio between TP and the sum of TP and FN (Eq. 1); while precision indicates the correct classifications among all classified as positive by the model, that is, the ratio between TP and the sum of TP and FP (Eq. 2).

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

The f1-score metric, used in the feature selection step, is defined as the harmonic mean between precision and recall, as presented in Eq. 3. Note that, if $TP = 0$, all positive samples are misclassified, and if $FP = FN = 0$, there is a perfect classification.

$$f1\text{-}score = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

The f1-macro average (f1-macro) is a variant of the f1-score, composed of the average of the f1-score of the positive class and the f1-score of the negative class (Eq. 4), in which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1

the impact on the final score increases since it is necessary to hit in both classes for good punctuation.

$$f1\text{-}macro = \frac{1}{m}\sum_{i=1}^{m} f1\text{-}score_i \quad (4)$$

3. Materials and Methods

To benchmark the machine learning models, we followed the methodology presented in Figure 1 with the goal to select the best model to assist the TB prognosis. The methodology adopted for this work includes preprocessing of the database; applying the feature selection algorithm in order to reduce the dimensionality of the database; training the models using an imbalanced database and a balanced database; applying randomized search technique to find the best hyper-parameters for the models; usage of statistical technique to determine whether models have similar distributions; finding the best models and creation of an ensemble model; usage of statistical technique to compare the best models; and finally, evaluation of the models through tests.

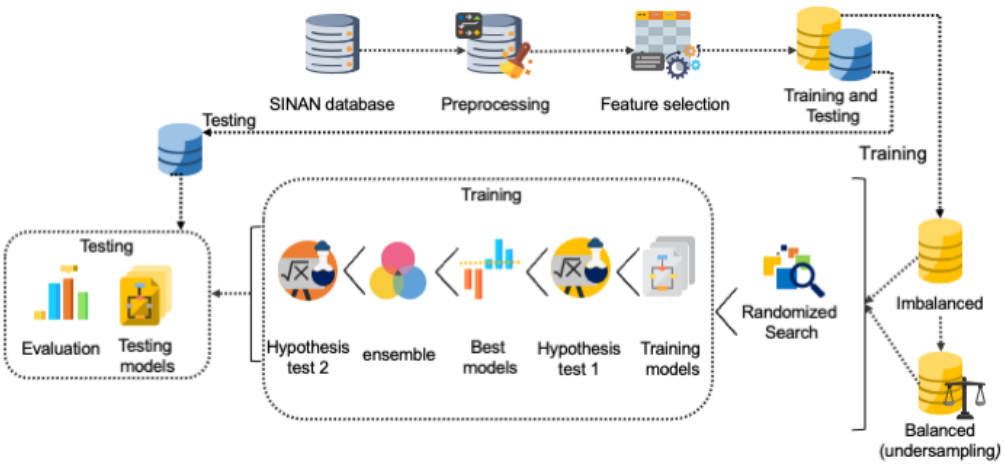


Figure 1. Methodology used to benchmark machine learning models.

The SINAN database contains records of patients related to diseases present on the Brazilian national list of compulsory reporting diseases. In this work, we use records related to the state of Amazonas from 2007 to 2018 of patients who were diagnosed and treated for TB. The SINAN-TB database has 1,221 TB death records and 35,007 cured patient records, totalling 36,228 records and 130 attributes. The description of all attributes can be consulted through the SINAN data dictionaries [48].

The preprocessing is performed as per [49], in order to clean the database for the model training and testing. After the preprocessing, the database resulted in 24,015 records with 38 features, being 1,139 related to the TB death class and 22,876 related to the TB cure class.

The feature selection technique called SFS was used to select the most representative features in the database, and then reduce the dimensionality of the data. In this work, 17 features were selected for each of the nine machine learning model, as per [9] that used the same SINAN-TB database and selected features by a specialist.

As the preprocessed database is imbalanced (1,221 TB death and 35,007 healthy patient) as the original one, two scenarios are designed for experiments and evaluations: using the original imbalanced database and using a balanced-version of the original database. To create the balanced database, the random under-sampling technique is applied and the balanced database is composed of 1,139 TB death and 1,139 healthy patient.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1

For both scenarios, the randomized search hyper-parameter optimiser is applied using the parameters e configurations available in the sci-kit learn library for Python[1].

Having selected the hyper-parameter configuration of each model, the models are trained using k-fold cross-validation ($k = 30$) and the average of the f1-macro metric is calculated.

The Wilcoxon hypothesis test is performed to eliminate models with similar distributions and know the best models (f1-macro greater than or equal to 80%). By eliminating models with similar distributions or with a performance below than 80%, the goal is to increase diversity for the ensemble model. Therefore, an ensemble model is built with the best models in order to improve the classification performance. Finally, given the best models, the test is performed 30 times and the f1-macro average is calculated for evaluation.

4. Results

4.1. Preprocessing and feature selection

As described in section 3, after applying the data preprocessing steps, the database's feature set was left with 38 features. In [9], Rocha et al. used the same SINAN-TB database and 17 features were selected by a specialist to predict the TB prognosis. In our work, for the application of the feature selection techniques, the same number of features was defined. We execute the feature selection techniques under k-fold cross-validation ($k = 5$), using the 9 different models.

Table 1 presents the 17 features that were selected by the SFS technique for each machine model. It is possible to notice that the models KNN (Figure 2c), NB (Figure 2e), SVM (Figure 2f) and RF (Figure 2h) presented a better f1-score with less than 17 features. However, in order to compare the models as per [9], the models were trained, validated and tested with 17 features. The features "DIAS" (days of hospitalization on which the patient spent treatment) and "BACILOSC_6" (result of sputum smear microscopy for bacillus alcohol resistant) were selected by all models. On the other hand, the features "OUTRAS" (other diseases and conditions) and "BACILOSC_O" (bacilloscopy of other material) were selected by only one model.

Table 1: Features selected by the SFS technique for each machine learning model

LR	LDA	KNN	DT	NB	SVM	GB	RF	MLP
CS_SEXO	CS_SEXO	RAIOX_TORA	CS_RACA	CS_SEXO	CS_SEXO	CS_RACA	TRATAMENTO	AGRAVAIDS
CS_RACA	CS_RACA	FORMA	TRATAMENTO	CS_RACA	CS_RACA	TRATAMENTO	FORMA	AGRAVOUTRA
TRATAMENTO	TRATAMENTO	AGRAVAIDS	RAIOX_TORA	TRATAMENTO	TRATAMENTO	TESTE_TUBE	AGRAVAIDS	BACILOSC_O
RAIOX_TORA	RAIOX_TORA	AGRAVDIABE	TESTE_TUBE	RAIOX_TORA	AGRAVAIDS	AGRAVAIDS	AGRAVDIABE	CULTURA_ES
FORMA	FORMA	AGRAVDOENC	FORMA	TESTE_TUBE	AGRAVALCOO	ETAMBUTOL	AGRAVDOENC	RIFAMPICIN
AGRAVALCOO	AGRAVAIDS	RIFAMPICIN	RIFAMPICIN	FORMA	AGRAVDIABE	ESTREPTOMI	AGRAVOUTRA	ISONIAZIDA
AGRAVDIABE	AGRAVALCOO	ISONIAZIDA	ISONIAZIDA	AGRAVAIDS	AGRAVDOENC	PIRAZINAMI	BACILOSC_E	ETAMBUTOL
AGRAVDOENC	AGRAVDIABE	ESTREPTOMI	ETAMBUTOL	AGRAVALCOO	AGRAVOUTRA	OUTRAS	CULTURA_ES	PIRAZINAMI
BACILOSC_1	AGRAVDOENC	PIRAZINAMI	PIRAZINAMI	AGRAVDIABE	BACILOSC_E	TRAT_SUPER	HIV	TRAT_SUPER
BACILOSC_2	AGRAVOUTRA	BACILOSC_3	BACILOSC_2	AGRAVDOENC	HIV	BACILOSC_1	BACILOSC_1	BACILOSC_2
BACILOSC_3	BACILOSC_E	BACILOSC_4	BACILOSC_3	AGRAVOUTRA	BACILOSC_3	BACILOSC_2	BACILOSC_2	BACILOSC_4
BACILOSC_4	HIV	BACILOSC_5	BACILOSC_4	BACILOSC_E	BACILOSC_4	BACILOSC_4	BACILOSC_4	BACILOSC_6
BACILOSC_5	RIFAMPICIN	BACILOSC_6	BACILOSC_5	CULTURA_ES	BACILOSC_5	BACILOSC_5	BACILOSC_5	TPUNINOT
BACILOSC_6	ETAMBUTOL	TPUNINOT	BACILOSC_6	HIV	BACILOSC_6	BACILOSC_6	BACILOSC_6	AGRAVDROGA
TPUNINOT	BACILOSC_6	AGRAVDROGA	AGRAVDROGA	BACILOSC_6	TPUNINOT	TPUNINOT	TPUNINOT	AGRAVTABAC
DIAS	DIAS	AGRAVTABAC	AGRAVTABAC	DIAS	DIAS	DIAS	DIAS	DIAS
IDADE	IDADE	DIAS	DIAS	IDADE	IDADE	IDADE	IDADE	IDADE

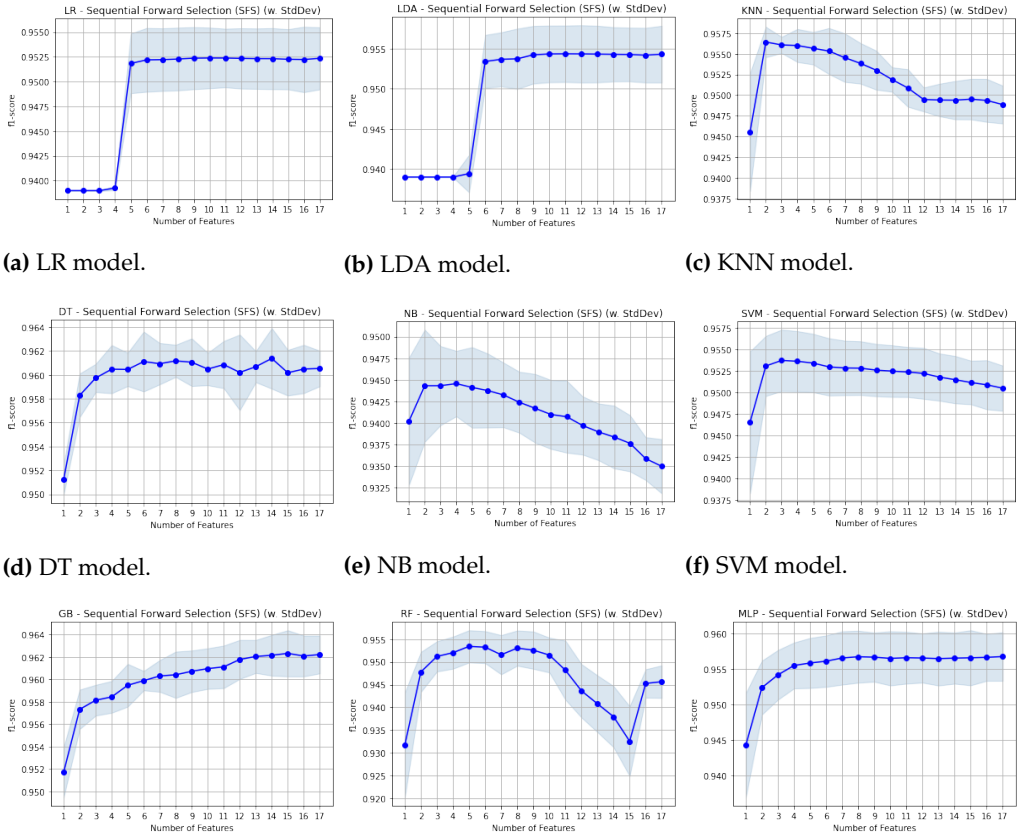
Figure 2 shows the f1-score results regarding the feature selection using SFS technique. The model that obtained the best mean f1-score was GB with 0.960064 ($\pm$0.002), and the worst was NB with 0.940909 ($\pm$0.003).

4.2. Results of the randomized search technique

Table 2 presents the best configuration for each model achieved by the randomized search technique for both scenarios (imbalanced and balanced database), considering the f1-macro as evaluation metric. These configurations were used to execute the training and testing of the models, presented next.

[1] https://scikit-learn.org/stable/supervised$_l$*earning.htmlsupervised* − *learning*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1



(a) LR model. **(b)** LDA model. **(c)** KNN model.

(d) DT model. **(e)** NB model. **(f)** SVM model.

(g) GB model. **(h)** RF model. **(i)** MLP model.

Figure 2. Result of f1-score metrics when applying the SFS feature selection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1

Table 2: Hyper-parameters configuration selected by the randomized search technique.

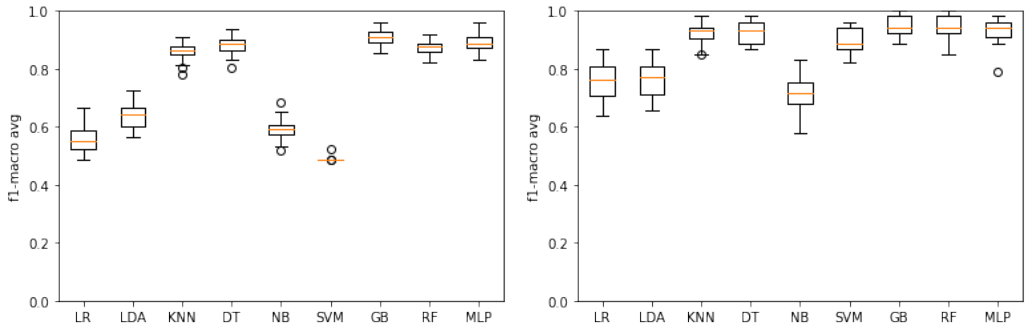
Model	Parameters	Randomized search using imbalanced database	Randomized search using balanced database
LR	Penalty	none	l1
	Solver	newton-cg	liblinear
	Multiclass	ovr	auto
LDA	Solver	svd	svd
	Shrinkage	None	None
	Priors	None	None
KNN	Weights	distance	distance
	Algorithm	ball_ree	ball_ree
	Leaf size	30	30
	Metric	minkowski	minkowski
	Parameter metric	None	None
	Number of jobs:	-1	-1
DT	Criterion	entropy	entropy
	Splitter	best	best
	Minimum samples split	3	4
	Minimum samples leaf	5	4
	Maximum features	sqrt	log2
SVM	Kernel	rbf	rbf
	Gamma	scale	scale
GB	Loss	exponential	exponential
	Criterion	friedman_mse	friedman_mse
	Number of estimators	300	300
	Minimum samples split	3	3
	Minimum samples leaf	4	4
	Maximum depth	9	9
	Maximum feature	log2	log2
RF	Criterion	entropy	entropy
	Number of estimators	200	200
	Minimum samples split	2	2
	Minimum samples leaf	1	1
	Maximum depth	6	6
	Maximum feature	log2	log2
	Maximum samples leaf	4	4
	Bootstrap	False	False
	OOB Score	False	False
	Weight class	balanced	balanced
MLP	Hidden layers	2	2
	Neurons in each layer	20	20
	Activation functions	logistic	logistic
	Solver	adam	adam
	Learning rate	invscaling	invscaling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1

As can be noted, hyper-parameters can change when using imbalanced and balanced databases. SVM, GB, RF and MLP models kept the same hyper-parameters configuration in both cases. For more details about the parameters and configurations, please refer the scikit-learn library.

4.3. Model training and validation

Figure 3a presents the results of the models' training regarding the f1-macro metric when using the imbalanced database. The model that obtained the best mean f1-macro was GB with 0.910506 (±0.024), and the worst was SVM with 0.488892 (±0.006). Figure 3b presents the results of the models' training regarding the f1-macro metric when using the balanced database. The model that obtained the best mean f1-macro was GB with 0.948488 (±0.029), and the worst was NB with 0.718603 (±0.059).



(a) Imbalanced database. **(b)** Balanced database.

Figure 3. f1-macro results of the machine learning models' training when using the (a) imbalanced and (b) balanced database.

Given the f1-macro results, the Wilcoxon test was applied in order to identify the models with similar distributions and discard the models with the lowest results.

When using the imbalanced database, KNN and DT; RF and MLP models presented similar distributions, and then KNN and DT were discarded. LR, LDA, NB and SVM models were discarded due the lowest results. Therefore, RF, GB and MLP models were selected to compose the ensemble model when considering the imbalanced database. Figure 4a presents the results of such models regarding the f1-macro when using the imbalanced database.

Regarding the balanced database, these following models presented similar distribution: LR and LDA; KNN, DT and MLP; GB and RF. In this case, LR, KNN, DT, RF and MLP models were discarded. LDA model were discarded due its low f1-macro. Therefore, two models were selected to compose the ensemble model in this case: SVM and GB. Figure 4b presents the results when using the balanced database.

Again, the Wilcoxon hypothesis test was executed. For the imbalanced database, no model has a similar distribution, so all models remain for the testing step. The ensemble was the best model (f1-score mean equals to 0.918638 (±0.024)).

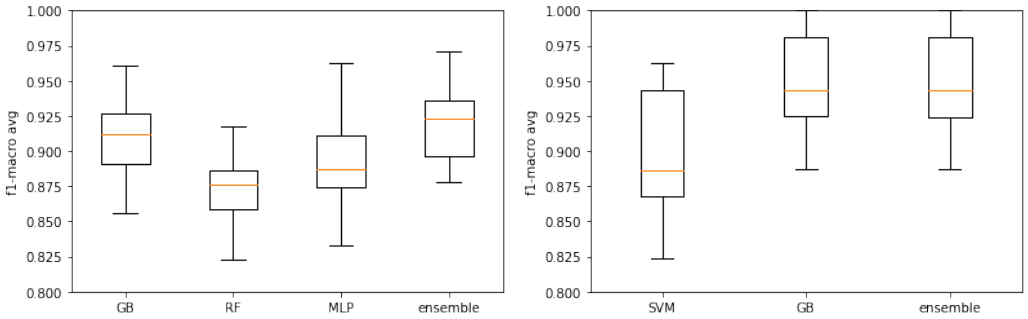
For the balanced database, ensemble model was discarded, and SVM and GB were the best models (f1-macro mean equals to 0.899186 (±0.043) and 0.948488 (±0.030), respectively). All results derived from the training step for the metric f1-macro are described in Table 3.

4.4. Testing the models

Given the models that presented the best performance during the training step, we test them using part of the database that were not used during the training. Figure 5 presents the f1-macro regarding the test step.

Ensemble model achieved the best mean f1-macro, 0.917389 (±0.001), using the imbalanced database 5a; and RF presented the worst result, 0.866524 (±0.002). For the balanced

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 March 2021 doi:10.20944/preprints202103.0284.v1



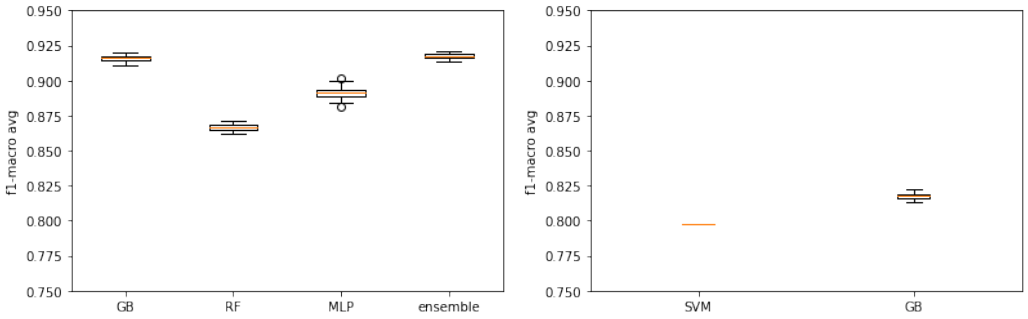
(a) Imbalanced database **(b)** Balanced database

Figure 4. Training of ensemble model and its respective models using the (a) imbalanced and (b) balanced database.

Table 3: Results of f1-macro metric regarding the models' training.

Models	Imbalanced database	Balanced database
LR	0.555333 (±0.042)	0.757205 (±0.064)
LDA	0.642187 (±0.043)	0.762497 (±0.061)
KNN	0.860864 (±0.030)	0.926506 (±0.035)
DT	0.882351 (±0.028)	0.930177 (±0.036)
NB	0.590700 (±0.033)	0.718603 (±0.060)
SVM	0.488892 (±0.006)	0.899186 (±0.043)
GB	0.910506 (±0.023)	**0.948488 (±0.030)**
RF	0.871556 (±0.023)	0.945916 (±0.033)
MLP	0.892978 (±0.028)	0.934526 (±0.041)
Ensemble	**0.918638 (±0.024)**	0.945942 (±0.031)

database 5b, the best mean f1-macro was the GB model with 0.817255 (±0.002), and SVM was the worst, 0.797650 (±0.000).



(a) Imbalanced database **(b)** Balanced database

Figure 5. Testing of models using the (a) imbalanced and (b) balanced database.

It is possible to observe the usage of imbalanced and balanced databases impact on the models' performance during the training phase (Figure 3a); and in general, models trained with the balanced database achieved superior results (Figure 3b). However, when the models were subjected to test (Figure 5), the ensemble model (composed of the RF, GB and MLP models) using the imbalanced database presented the best performance regarding f1-macro metric when compared against GB model using the balanced database. All results derived from the testing step for the metric f1-macro are described in Table 4.

This result shows that submitting a database to a random balancing technique, such as under-sampling, can result in loss of performance for machine learning models, as important data may be lost during the sampling.

Table 4: Results of f1-macro metric regarding the models' testing.

Models	Imbalanced database	Balanced database
SVM	-	0.797650 (±0.000)
GB	0.916041 (±0.002)	**0.817255 (±0.002)**
RF	0.866524 (±0.002)	-
MLP	0.891276 (±0.004)	-
Ensemble	**0.917389 (±0.001)**	-

5. Conclusions

Despite the existence of several works on TB diagnosis, there is a lack on TB prognostic, which is a critical factor in effective TB treatment. In this paper, we addressed an important gap in the literature by performing a benchmarking of several machine learning models to assist the TB prognosis. An ensemble model was also proposed, considering heterogeneous classifiers and presented the best performance.

Regarding the performance of the machine learning models with the imbalanced database, the ensemble composed of RF, GB and MLP models presented the best f1-macro, 0.917389 (±0.001). When using the balanced database, the GB model obtained the best f1-macro, 0.817255 (±0.002). Based on these results, we can state that the database preprocessing impacted directly on the performance of the models, but further investigation on other metrics would bring more insights about models performance.

As future work, we plan to develop a framework composed of the best models to assist health professionals in the treatment of TB. This framework will provide a mobile application, in which users will be able to make use of our models in order to support their decision regarding the treatment after receiving a patient diagnosed with TB.

References

1. Pai, M.; Behr, M.; Dowdy, D.; Dheda, K.; Divangahi, M.; Boehme, C.; Raviglione, M. Tuberculosis. Nature Reviews Disease Primers, 2, 16076, 2016.
2. WHO. Global Tuberculosis Report 2020. https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf. Accessed: 2021-01-25.
3. Tuberculosis profile: Brazil. https://worldhealthorg.shinyapps.io/tb_profiles\/?_inputs_&lan=%22EN%22&iso2=%22BR%22. Accessed: 2020-09-25.
4. WHO. Country profiles FOR 30 HIGH TB BURDEN COUNTRIES. https://www.who.int/tb/publications/global_report/tb19_Report_country_profiles_15October2019.pdf?ua=1. Accessed: 2020-09-29.
5. Sistema Único de Saúde (SUS): estrutura, princípios e como funciona. https://antigo.saude.gov.br/sistema-unico-de-saude. Accessed: 2021-01-25.
6. Brasil é único com 'SUS' entre países com mais de 200 milhões de habitantes. https://www1.folha.uol.com.br/cotidiano/2019/10/brasil-e-unico-com-sus-entre-paises-com-mais-de-200-milhoes-de-habitantes.shtml. Accessed: 2021-01-28.
7. Brazil's Sistema Único da Saúde (SUS): Caught in the Cross Fire. https://www.csis.org/blogs/smart-global-health/brazils-sistema-unico-da-saude-sus-caught-cross-fire. Accessed: 2021-01-25.
8. Martins, V.d.O.; de Miranda, C.V. Diagnóstico e Tratamento Medicamentoso Em Casos de Tuberculose Pulmonar: Revisão de Literatura. *Revista Saúde Multidisciplinar* **2020**, *7*.
9. Rocha, E.d.S. DEEPTUB: Plataforma para predição de morte por tuberculose baseado em modelos de Deep Learning utilizando dados demográficos, clínicos e laboratoriais. Dissertação de Mestrado, Universidade de Pernambuco, 2020.
10. Lakhani, P.; Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **2017**, *284*, 574–582.
11. Rajaraman, S.; Candemir, S.; Xue, Z.; Alderson, P.O.; Kohli, M.; Abuya, J.; Thoma, G.R.; Antani, S. A novel stacked generalization of models for improved TB detection in chest radiographs. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018, pp. 718–721.

12. Hooda, R.; Sofat, S.; Kaur, S.; Mittal, A.; Meriaudeau, F. Deep-learning: A potential method for tuberculosis detection using chest radiography. 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, 2017, pp. 497–502.
13. Sethi, K.; Parmar, V.; Suri, M. Low-Power Hardware-Based Deep-Learning Diagnostics Support Case Study. 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2018, pp. 1–4.
14. Kant, S.; Srivastava, M.M. Towards automated tuberculosis detection using deep learning. 2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2018, pp. 1250–1253.
15. Carneiro, G.; Oakden-Rayner, L.; Bradley, A.P.; Nascimento, J.; Palmer, L. Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, 2017, pp. 130–134.
16. Song, Q.; Zheng, Y.J.; Xue, Y.; Sheng, W.G.; Zhao, M.R. An evolutionary deep neural network for predicting morbidity of gastrointestinal infections by food contamination. *Neurocomputing* **2017**, *226*, 16–22.
17. Lee, C.K.; Hofer, I.; Gabel, E.; Baldi, P.; Cannesson, M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology: The Journal of the American Society of Anesthesiologists* **2018**, *129*, 649–662.
18. Bora, R.M.; Chaudhari, S.N.; Mene, S.P. A Review of Ensemble Based Classification and Clustering in Machine Learning.
19. García-Gil, D.; Holmberg, J.; García, S.; Xiong, N.; Herrera, F. Smart Data based Ensemble for Imbalanced Big Data Classification. *arXiv preprint arXiv:2001.05759* **2020**.
20. Yang, K.; Yu, Z.; Wen, X.; Cao, W.; Chen, C.P.; Wong, H.S.; You, J. Hybrid Classifier Ensemble for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems* **2019**, *31*, 1387–1400.
21. Kira, K.; Rendell, L.A. A practical approach to feature selection. In *Machine Learning Proceedings 1992*; Elsevier, 1992; pp. 249–256.
22. Marcano-Cedeno, A.; Quintanilla-Domínguez, J.; Cortina-Januchs, M.; Andina, D. Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. IECON 2010-36th annual conference on IEEE industrial electronics society. IEEE, 2010, pp. 2845–2850.
23. ALAKUŞ, T.B.; TÜRKOĞLU, İ. Feature selection with sequential forward selection algorithm from emotion estimation based on EEG signals. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi* **2019**, *23*, 1096–1105.
24. Sammut, C.; Webb, G.I. *Encyclopedia of machine learning and data mining*; Springer, 2017.
25. Das, K.; Behera, R.N. A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering* **2017**, *5*, 1301–1309.
26. Callahan, A.; Shah, N.H. Machine learning in healthcare. In *Key Advances in Clinical Informatics*; Elsevier, 2017; pp. 279–291.
27. Bonte, C.; Vercauteren, F. Privacy-preserving logistic regression training. *BMC medical genomics* **2018**, *11*, 86.
28. Menard, S. *Applied logistic regression analysis*; Vol. 106, Sage, 2002.
29. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear discriminant analysis. In *Robust data mining*; Springer, 2013; pp. 27–33.
30. Balakrishnama, S.; Ganapathiraju, A. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* **1998**, *18*, 1–8.
31. Basha, S.M.; Rajput, D.S. Survey on Evaluating the Performance of Machine Learning Algorithms: Past Contributions and Future Roadmap. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems*; Elsevier, 2019; pp. 153–164.
32. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. KNN model-based approach in classification. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, 2003, pp. 986–996.
33. Talita, A.; Nataza, O.; Rustam, Z. Naïve Bayes Classifier and Particle Swarm Optimization Feature Selection Method for Classifying Intrusion Detection System Dataset. Journal of Physics: Conference Series. IOP Publishing, 2021, Vol. 1752, p. 012021.
34. Rukmawan, S.; Aszhari, F.; Rustam, Z.; Pandelaki, J. Cerebral Infarction Classification Using the K-Nearest Neighbor and Naive Bayes Classifier. Journal of Physics: Conference Series. IOP Publishing, 2021, Vol. 1752, p. 012045.

35. Rish, I.; others. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3, pp. 41–46.
36. da Silva, L.A.; Peres, S.M.; Boscarioli, C. *Introdução à mineração de dados: com aplicações em R*; Elsevier Brasil, 2017.
37. Bordoloi, D.J.; Tiwari, R. Optimum multi-fault classification of gears with integration of evolutionary and SVM algorithms. *Mechanism and Machine Theory* **2014**, *73*, 49–60.
38. Yao, Y.; Liu, Y.; Yu, Y.; Xu, H.; Lv, W.; Li, Z.; Chen, X. K-SVM: An Effective SVM Algorithm Based on K-means Clustering. *JCP* **2013**, *8*, 2632–2639.
39. Lu, H.; Karimireddy, S.P.; Ponomareva, N.; Mirrokni, V. Accelerating Gradient Boosting Machines. International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 516–526.
40. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* **2013**, *7*, 21.
41. Gomes, H.M.; Bifet, A.; Read, J.; Barddal, J.P.; Enembreck, F.; Pfharinger, B.; Holmes, G.; Abdessalem, T. Adaptive random forests for evolving data stream classification. *Machine Learning* **2017**, *106*, 1469–1495.
42. Zanaty, E. Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal* **2012**, *13*, 177–183.
43. Zhou, Z.H. *Ensemble methods: foundations and algorithms*; CRC press, 2012.
44. Liashchynskyi, P.; Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. *arXiv preprint arXiv:1912.06059* **2019**.
45. Woolson, R. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* **2007**, pp. 1–3.
46. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* **2020**, *21*, 1–13.
47. Lipton, Z.C.; Elkan, C.; Naryanaswamy, B. Optimal thresholding of classifiers to maximize F1 measure. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2014, pp. 225–239.
48. Dicionário de dados - SINAN NET - Versão 5.0. http://portalsinan.saude.gov.br/images/documentos/Agravos/Tuberculose/DICI_DADOS_NET_Tuberculose_23_07_2020.pdf. Accessed: 2021-01-25.
49. Lino Ferreira da Silva Barros, M.H.; Sampaio, Vanderson, E.; Takako, P. Aplicação de técnicas de record linkage e feature selection para análise e seleção de características em uma base de dados integrada do SINAN, 2020.

 doi:10.20944/preprints202103.0284.v1