Article

Data Fusion Using Medical Records and Clinical Data to Support TB Diagnosis

Andrés F. Romero-Gómez , Alvaro D. Orjuela-Cañón * , Andrés L. Jutinico , Carlos E. Awad , Erika Vergara , María A. Palencia

Posted Date: 19 June 2024

doi: 10.20944/preprints202406.1316.v1

Keywords: artificial intelligence; tuberculosis diagnosis; data fusion

Article

Data Fusion Using Medical Records and Clinical Data to Support TB Diagnosis

Andrés F. Romero-Gómez [1,†], Alvaro D. Orjuela-Cañón [2,†], Andres L. Jutinico [3,†], Carlos E. Awad [4,†], Erika Vergara [5,†] and Maria A. Palencia [4,]

[1] Fundación Santa Fe de Bogotá; andres.romero@fsfb.org.co
[3] School of Medicine and Health Sciences, Universidad del Rosario; alvaro.orjuela@urosario.edu.co
[3] Biomedical Engineering, Universidad Antonio Narino; ajutinico@uan.edu.co
[4] Subred Integrada de Servicios de Salud Centro Oriente, Bogota D.C., Colombia; carlosawad@gmail.com
[5] Hospital Universitario Nacional; evergarav@hun.edu.co
[*] Correspondence: alvaro.orjuela@urosario.edu.co; Tel.: +57 1 2970200 - 3479
[†] These authors contributed equally to this work.

Abstract: Tuberculosis (TB) is an infectious disease declared a global emergency by the World Health Organization and continues as one of the world's top ten causes of death. TB diagnosis is more critical in developing countries where demanded infrastructure for detection, and treatment complicates the efforts against the disease. These aspects related to limited resources are significant, especially in areas away from the main cities, with few mechanisms to make a timely diagnosis that contributes to successfully addressing the possible patients. Artificial intelligence has begun to be essential in providing additional strategies in the diagnosis processes for health professionals' support. This paper uses natural language processing (NLP) and machine learning (ML) techniques to create models that can supply TB diagnosis when the needed infrastructure is unavailable. Two different sources were explored: text extracted from electronic medical records (EMR) and patient clinical data (CD). Four proposals using five different machine learning models were implemented. The first two models employed ML and each data source independently. Then, two additional approaches developed a data fusion from both sources. This strategy's employment was analyzed with physicians according to their pertinence in the process and understanding of the EMR. Finally, the results of the data fusion were compared to each source, obtaining better performance at using only the CD, where an area under the ROC curve of 69.92.3% was obtained. However, the advantage of analyzing physician's reports is the availability of this information contrasted to clinical-specific data, which can be more useful in places far from the main cities without enough basic structure for its obtaining.

Keywords: artificial intelligence; tuberculosis diagnosis; data fusion

1. Introduction

Tuberculosis (TB) is caused by *Mycobacterium tuberculosis* which is spread from person to person through the air, which makes it highly contagious. According to this, the World Health Organization (WHO) maintains that TB is among in the top ten causes of death worldwide. Moreover, until 2020, it was the leading cause of death caused by a single infectious agent, now overshadowed by COVID-19 [1,2]. In developing countries, the situation is worse due to almost 90% of the people with active TB are concentrated in 30 of these countries each year. Different efforts led by WHO, such as the End TB initiative, have been proposed, but it has been difficult to reduce the incidence rates (newly diagnosed cases), which had an increment of 3.6% between 2020 and 2021, due to the impact of the COVID-19 pandemic. In addition, it was estimated that 10.6 million people fell ill with TB in 2021, representing 4.5% more than in 2020 [3].

One out of four people is estimated to be infected with latent TB worldwide. This means that the mycobacterium lives in the host without developing and transmitting the disease. Furthermore, people with this latent TB can develop active TB with a $5-10\%$ more risk associated with compromised immune systems such as HIV, diabetes, malnutrition, tobacco, or homeless conditions [4], [5]. TB can affect any organ of the body but mainly attacks the lungs, known as pulmonary TB (PTB), with symptoms such as severe cough lasting more than three weeks, chest pain, and coughing up blood or sputum [6], [7]. Other forms of TB as recognized as extra-pulmonary TB, where the most common is pleural TB, and the more lethal is meningeal TB.

The present work analyzed a case with information from a developing country such as Colombia. This country has reported an incidence rate of 21.83 in 2021 and an increase to 27.19 in 2022 per 100000 population, registering 17341 new reported cases in the last year for PTB [3,8]. There, it is straightforward to identify some problems related to the inequality of resources regarding public health. There are regions or states in the national territory with higher PTB incidence rates, such as Amazonas or Antioquia, where the rural conditions established values of 75.55 and 46.72, respectively. This last state had 19.8% of the total PTB cases in Colombia [1,8,9]. These farming areas with precarious health systems are far from the main cities, as the capital, where the incidence rate reached 16.02. It is necessary to contribute to proposing alternative strategies that allow better disease management and identification and begin the antiPTB treatment as soon as possible.

In most cases, PTB can be treated aggressively using four drugs for six months or more in cases of drug-resistant TB [10,11]. The Colombian health system has a protocol for these treatments [12], but in some regions, the diagnostic methods are not fully accessible, presenting fatal consequences. The same standard has PTB detection and reporting guidelines in the national health system [12]. Besides, the diagnosis must be elaborated by microbiological confirmation of the presence of the mycobacteria in the sputum of suspicious patients. For this, there are three types of tests: smear microscopy, molecular tests, and culture. However, the protocol established that health professionals can initiate the treatment based on clinical analysis of the patient, even without bacteriological confirmation, to prevent the disease from spreading and to avoid its progress in the patient.

Smear microscopy is the simplest and least expensive test, with results that can take short periods. Despite this, its sensitivity is low, between 40% and 60%, depending on the quality of the sample. Molecular tests have a higher sensitivity, more than 90%, and results can be delivered within hours. The disadvantage of this method is the specialized requirements related to equipment and professionals, which can become more expensive than smear microscopy. Finally, cultures are the most potent methods, with high sensitivity and specificity, demanding skilled personnel and costly infrastructure, and the duration of the results can take two to three weeks at best [12,13]. Each test has its advantages and disadvantages. Depending on the availability of the tests, a patient may have one, two, or all three tests performed. Nevertheless, the time and costs associated with each test constitute a gap in its accessibility in some places, so it is necessary to create new low-cost and rapid technologies that can be used to support health professionals in diagnosing the disease [14].

Despite the current technologies and the protocol to diagnose the PTB, some locations in the country need more conditions to develop this procedure. Sometimes health professionals hold traditional instruments to do their work and laboratories or sophisticated supplies can only be performed on time. Based on previous strategies reported by the same team of researchers, a search for alternatives for this scenario is related in [15]. In the present case, an extension was based on a reported text that medical staff accomplished in the patient consultation, introducing a description with medical findings and terminology, which holds in the traditional Colombian health systems and can be analyzed as an additional data source in the diagnosis process.

In recent years, Artificial Intelligence (AI) has been used in medicine to support decisions [16,17]. The findings are supported by computer systems on which health professionals can rely to do a better job. These systems can process large volumes of data and use that learning for a specific task [18,19]. Among the advantages of these tools is that they are low-cost, and front-line personnel can use them, making them useful in situations where conventional methods are unavailable [20]. Different applications of using AI in health problems can be seen in support of cancer detection [21], COVID-19 diagnosis and treatment, and drug discovery through deep neural networks implementation [22–24].

In the case of TB, [25] and [26] show how an artificial neural network (ANN) can be trained to diagnose TB using clinical data (CD), and in [26] and [27] variants of ANN for clustering are used to determine three risk groups (high, medium, and low risk) of the population concerning TB, showing good results.

Natural Language Processing (NLP) is a branch of AI that allows the analysis of texts that are not necessarily written in a structured language and can be used to design support tools. For example, NLP has been used to build AI systems that help in tasks such as searching for relevant information [28], determining eligibility and tracking patients [29,30], or diagnosing diseases [31], generally using information contained in the electronic medical records (EMR) such as clinical, laboratory or image data. For constructing these NLP systems, it has been found that the best-performing models are those that learn the rules of the data, i.e., those that use machine learning (ML) techniques [32]. From 2021 to now, the popularity of large language models (LLM) has increased, motivating the use of generative models and proposals for text processing. Difficulties presented by these approaches are associated with the specific dataset employed to develop the model and the manual annotations made by expert medical professionals [33,34]. Examples of this can be seen in diabetes disease problem for evaluating the risks related to food intake [35], a specific TB case with a comparison between NLP labels based on image and the TB experts screening on chest RX [36], among other applications for data fusion in the medical context [37]. An important difference from those previous works, we explored the use of the NLP representation according to available information for training the models in a limited scenario with little data. This represents the context of some developing countries, where there are challenges related to systematic health data acquisition, available data, and the health human resources component included in model development.

The present work is framed within that project and seeks to develop models based on computational intelligence to allow health personnel to make better decisions about suspected PTB patients. The data was collected from 151 patients, consisting of clinical data such as HIV status, location inside the city, and sex. In addition, physicians report (PR) at the moment of patient consultation as a routine practice. With these two sources of information, four schemes were proposed to create models that predict PTB status from data. This data fusion proposal explores information provided by CD, PR, or both sources in the PTB diagnosis support process. For this, two schemes merge the information from CD and PR, and the other two use them separately. In all of them, the use of five ML algorithms was explored: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and ANN. In the case of PR, it is necessary to employ NLP techniques to represent the texts in numerical form so that the ML algorithms can learn from this representation. The proposed scenario can contribute to training models employed in tasks associated with diagnosing PTB without the main cities' evolved infrastructure.

2. Materials and Methods

Figure 1 shows a general scheme of the methodology implemented. First, data were acquired at the Hospital Santa Clara in Bogota D.C., Colombia. Then, a preprocessing and cleaning process was implemented before extracting the features representing the data, which is the input of the algorithms. For the ML models, four types of schemes that use five ML algorithms were proposed, and a repeated through stratified k folds cross-validation technique was performed to evaluate the performance of the models. Each step of Figure 1 will be explained in detail in this section.

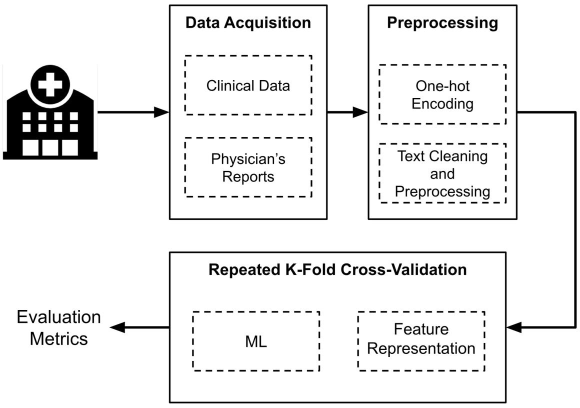

Figure 1. General scheme of the methodology implemented.

Dataset

The data were acquired during the period from 2017 to 2019 at the Hospital Santa Clara in Bogota D.C., Colombia, associated with the Middle-East Integrated Subnetwork of Health Services of the city (from Spanish *Subred Integrada de Servicios de Salud Centro Oriente*). The subnetwork ethics committee approved the study by act No 316 of May 24^{th}, 2021. Approval was established based on anonymization of data with only population-related variables that posed no risks to subjects. Informed consent was not required because all data were retrospective and anonymous.

Data variables associated with CD were extracted from registers from the institutional PTB program. PR was obtained from EMR, which is standardized according to the record acquisition system of the hospital. Data belong to clinically suspected subjects of having PTB, and microbiological tests were performed on these patients to confirm the disease. These tests allowed us to determine 116 confirmed PTB cases and 35 patients without the disease from an initial dataset with 233 individual registers. The difference was not considered because of unavailable PTB confirmation or PR information.

For PR extraction, when usable, it was decided to employ dates related to diagnostic tests and treatment beginning. Temporal locating of the PR was taken into account due to the information holding that suggests PTB. Thus, five PR were extracted 30 days before to these dates and analyzed and summarized because not all the information in the EMR can be associated with PTB, but the patient's entire medical history.

In the case of the CD, data were collected by the physicians involved in the PTB institutional program and members of the researchers' team, including patient's information such as *i)* Bogota's inside location, *ii)* sex, *iii)* human immunodeficiency (HIV) status, *iv)* antiretroviral treatment (ART) status, *v)* type of population associated with risk groups such as homeless, migrants, displaced persons, or belonging to indigenous people. These variables were considered according to suggestions from physicians and their essential codification by the binary representation [38]. Table 1 visualizes the variables and possible values.

Table 1. Variables used for the CD source data.

Variable	Values
Sex	Male Female
Type of Population	Homeless Native Exile Immigrant Prison Violence Victim Other
City Location	Antonio Nariño Barrios Unidos Bosa Chapinero Ciudad Bolivar Engativa Fontibón Kennedy La Candelaria Los Mártires Puente Aranda Rafael Uribe Uribe San Cristóbal Santa Fe Suba Teusaquillo Tunjuelito Usaquen Usme Out of Bogota City Unknown
HIV Status	Yes No Unknown
Antiretroviral Treatment Status	Yes No Unknown

Preprocessing

In the preprocessing and data cleaning, an exploratory analysis was performed to determine any errors in acquiring the data and to fix them if possible. The database of 151 patients was constructed considering this analysis, because there were cases of patients in which, for external reasons, the EMR was not available, or the dates recorded for the performance of the diagnostic tests did not correspond to those appearing in the EMR. It was also decided to organize the population of Bogota into five zones based on where they live in the north, south, east, west, and outside Bogota so that there is a significant number of patients per zone. The localities in each area correspond to those assigned by the city to each one of the four integrated subnetworks health services. This allows including indirect sociodemographic information.

On the other hand, the preprocessing of the PRs consists of cleaning up the text and removing Spanish stop words to reduce noise and ensure that the text representation (feature extraction) and the ML focus more on words and concepts that can provide relevant information regarding PTB. The stop words are meaningless words within a language, such as articles, pronouns, and prepositions, and in this case, units of physiological variables were included. Removing stop words is relevant in this work because these meaningless words do not provide information for the classification task.

Other words in the clinical reports indicate information and events unrelated to PTB [39]. However, the effect of these last terms is intended to be reduced with post-processing. Additionally, characters that were not letters of the Spanish alphabet were eliminated, including accents, punctuation marks, numbers, and other strange symbols. Also, double spaces and line breaks were eliminated. All this was done to remove from the text elements that are not useful and may confuse the algorithms. More information about the construction of the dataset from the EMR and a preliminary analysis of the content of the texts can be found in [39]. For practical issues, the patient document is called the union of the five PR preprocessed and cleaned for each patient.

Feature Representation

The CD source data was established from nominal variables represented by numerical vectors. For this, the variables were one-hot encoded, employing binary unique values associated to the original variable value. For example, the variable HIV has three values: positive, negative, and unknown. Therefore, three columns were created, representing each value, assigning one (1) binary value for the corresponding relation of existence or zero (0) in the absence of the value variable. This numerical representation allows ML algorithms to be trained from the categorical variables of the CD without a specific order or preference.

For NLP applications, computers receive language in the form of text coming from the patients' documents. These texts need to be encoded into a numerical representation so that ML algorithms can learn from the data. For this, two modes of representing the text of the patients' documents were explored, in which each document has a vector representing its content.

Term Frequency - Inverse Document Frequency: The first method uses the measure Term Frequency- Inverse Document Frequency (TF-IDF) [40]. This metric helps highlight words that appear in a document but penalizes them if they appear in several documents. The information of the word is related inversely to the number of times that appears in the documents, due to the low relevance when the classifying of the documents is done. Equation (1) shows how IDF was calculated for each term (t), where n is the number of documents, and $DF(t)$ is the number of documents in the document set that contain the term (t). The set of all the terms that appear in the documents is known as the vocabulary.

$$IDF(t) = log\left(\frac{1+n}{1+DF(t)}\right) + 1, \tag{1}$$

After obtaining the IDF for all the terms, the TF-IDF is computed by multiplying each item (TF) occurrence in a document with the corresponding IDF. Thus, each document has a vector with the TF-IDF value for each term in the set. In addition, an L2 normalization was applied for each vector, so the sum of squares of the vector elements is 1.

On the other hand, the terms for TF-IDF do not necessarily have to be single words, the use of *n-grams* can be implemented. *N-grams* are sub-sequences of elements, for example, a 2-gram is a sequence of two consecutive words that appear in a document. This strategy can capture the language structure and add context to the words. Different combinations of *n-grams* were explored to build the vocabulary: *i)* 1-grams (a single term), *ii)* 1-grams and 2-grams combination, *iii)* 2-grams, *iv)* 1-grams, 2-grams, and 3-grams combination, *v)* 3-grams. As the number of terms in the vocabulary can be large, the size of the vectors was limited to a range of $DF(t)$, the maximum and minimum values of that range were explored during the process, and the vocabulary size was also limited to 1000 terms at most. However, the length of the vectors remains large to apply ML with the number of samples available, so a dimensionality reduction was employed using truncated single value decomposition (SVD), where the number of components was also explored.

Version 1.0 of the *Scikit Learn* library was used to obtain the TF-IDF vector for each document [41]. The vector computation was based on parameters explained above, such as analyzer by words, *n-grams* from one to three grams, and a maximum number of features of 1000. According to the previous subsection, stop-words were not considered when a preprocessing was implemented based on the Spanish language.

Embeddings Representation: The second method employed embeddings representation. In the embeddings, each term is represented by a vector so that words with similar meanings, or used in a similar context, have a close vector representation. There are different ways to obtain embeddings, they can be obtained from a layer of an ML model by training in a specific task, or they can also be obtained in an unsupervised way from document statistics [42]. In the present work, Word2Vec model was used to build the embeddings [43]. There, two algorithms were applied, continuous bag-of-words (COBW), which predicts a word in the middle of a window using words around it. The second algorithm is skip-grams, the reverse of COBW one, which indicates the context word for a given target word based on unsupervised learning techniques. Both methods employ neural networks, layer, and weights concepts to predict the words and to obtain the embeddings.

Finally, there is a vector that represents the terms in the document, but a vector that represents each document is needed to use it to train classic ML algorithms. For this reason, a pooling step was added using two different operations, the first one by calculating the maximum between the elements of the vectors of each word and the other one by calculating the mean. Therefore, each document was represented by a vector containing either the maximum or the mean of the set of embeddings that represents each word.

Machine Learning Models

Before discussing the experimental design, it was necessary to understand the ML algorithms used briefly. As mentioned above, the five algorithms used are KNN, LR, SVM, ANN, and RF. These models were trained to do a binary classification between patients with confirmed TB and patients without TB. KNN is an algorithm that does not attempt to construct a general internal model, it simply stores instances of the training data, so it is an instance-based learning classifier. The classification was computed using the information of the k (a specified number) nearest neighbors. In this approach, the discrimination is based on the voting of the majority of the samples with the nearest neighbors or assigning weights according to the inverse proportional distance from the point to be classified [44].

The regression task developed by the LR model consists of optimizing a cost function. There, the problem is to find the probability of doing the best job based on a logistic regression function, modeling the probability for a binary output, which in the present case is the TB detection [41].

In the case of the ANN, different architectures can be employed, however, the current one consists of one hidden layer with a set of neurons. Each neuron performs a weighted sum of its inputs and applies an activation function on the output to add nonlinearity to the data. The weights of the weighted sum are learned from the training set and generalized for new inputs [45].

Another approach for classification can be implemented through the use of SVM. The objective here is to maximize the margin, defined as the distance between the separation hyperplane and the training labeled samples closest to this hyperplane [46]. These samples are called support vectors and contribute to the classification of two main classes: TB-detected and non-detected.

Finally, the RF model assembles a set of decision trees that are trained separately like a real forest. This training has the objective of maximizing the information. However, one of these trees usually needs to be more capable of solving a classification problem independently. What RF achieves is to put together a set of weak classifiers to improve the classification task [47]. The *Scikit Learn*library was used to implement all algorithms except ANN, which was implemented with Keras library and its version 2.7.0rc0 [48]. For each case, the hyperparameters or each approach were established through a grid search strategy, where different values were tested and those with better results were chosen.

Experimental Design

As mentioned above, five ML schemes were proposed to perform the classification. Figure 2 shows these schemes. Two approaches use PR and CD isolated and two fusion schemes that merge the two sources of information: type A and type B data fusions. The ML approaches were employed for the PR and CD-based and type A data fusion models. By contrast, in the type B data fusion models, where the ML-employed models were selected based on the results of PR and CD-based isolated approaches. This was because it is possible to know which are the best ML algorithms for the problem.

In the type B data fusion model, there is an aggregation layer with two different strategies. The first one is a simple majority vote of the four classifiers. In cases of a tie, the classifier selects the class by giving the highest priority to model one and the lowest priority to model four. The second strategy is called stacking classifier, where an ML algorithm is used as an aggregation layer. The LR algorithm was chosen for this task because its inputs are the four ML models' probability or the decision function (depending on the type of algorithm).

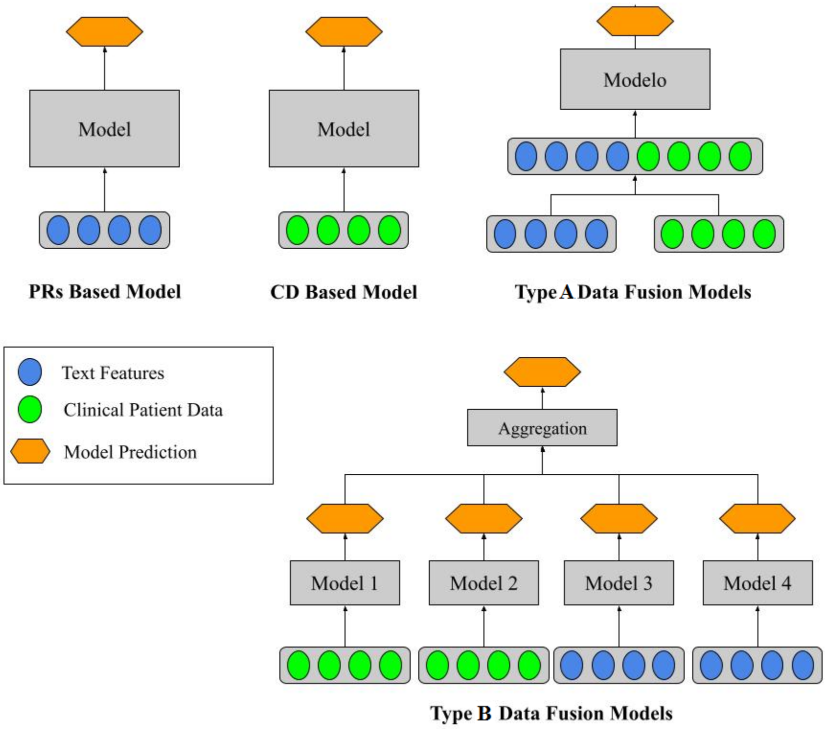

Figure 2. The four proposed ML classification schemes.

On the other hand, a different model was added from those mentioned above, which consists of a neural network as shown in Figure 3. This architecture was designed so that two networks would learn simultaneously from the PR and the CD in the learning process. Then, these layers are concatenated, and a new layer oversees the information fusion. The output layer has an output neuron with a sigmoid activation function that does the binary classification.

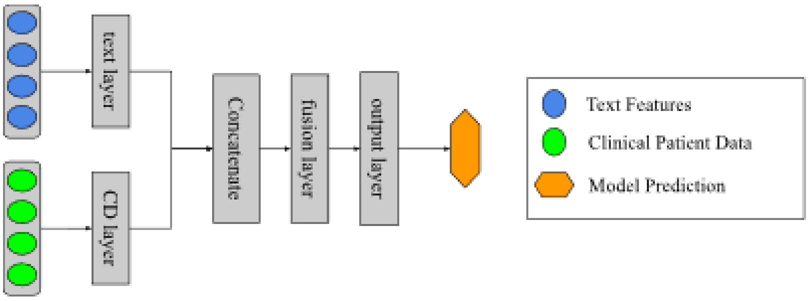

Figure 3. ANN used for information fusion (ANN Fusion).

Stratified *k*-fold cross-validation (SKCV) was used to select the best parameters for text representation with TF-IDF, embeddings, and the hyperparameters of the ML models, as shown in Figure 1. In SKCV, the data were separated into different partitions for training and validation of the approaches, with the objective of being different sets. Stratified means that the divisions preserve the percentage of samples for each class. Three folds were taken because of the database size, maintaining a considerable quantity in the validation set for the evaluation of models correctly. For statistical performance evaluation, the SKCV was repeated ten times, and the mean and standard deviation over the metrics of all repetitions were employed to evaluate and compare the results of the models.

The parameters explored to obtain the TF-IDF measure were the combinations of *n-grams* and the maximum and minimum values of $DF(t)$ to include a term in the vocabulary. The number of components for the dimensionality reduction was also explored with the TF-IDF calculated. Both, the calculation of the TF-IDF and the dimensionality reduction, were fitted by employing the training set, and then, applied to the training and testing sets. The methods were fitted with the training set because when predicting the diagnosis of a document, the TF-IDF method does not know the vocabulary within that document. In this way, the test documents simulate adding a new input well.

For the Word2Vec models obtention, the training set was needed to implement the neural networks of the COBW and skip-grams models. In the embedding models, parameters such as the size of the window and the minimum number of times a word appears to be part of the vocabulary were explored. Also, in the pooling step, the maximum and the mean operation were used. For both feature extraction methods, a vector that represents each document with a normalization where a transformation with values into a range between -1 and 1 was applied.

Furthermore, the hyperparameters of the ML algorithms were explored by employing random grid values in a heuristic mode until the best results were found. For models with a small size of hyperparameters, every single option was examined whenever the problem allowed it: LR, ANN, and the SVM kernel models. In the fusion models, the same methodology was also performed. However, this exploration was shorter because of the results for the PR and CD-based models. Also, it is important to note that a combination of the parameters of the text representation and the ML hyperparameters were considered in a coupled mode.

The metrics used to evaluate the results of the SKCV were sensitivity, specificity, and area under the curve (AUC), according to its pertinence in health systems [49]. These metrics were calculated for each of the testing sets and averaged. Subsequently, ten repetitions were performed to find the average and standard deviation over ten cross-validations using SKCV. The mean and standard deviation of these three metrics allowed us to evaluate the variability of the models. Additionally, due to the unbalanced dataset, the AUC was weighted with the number of samples of each class to have a metric that considers this unbalanced. Finally, the same training and test sets were always used to compare models' results.

Results and Discussion

Tables 3-5 show the results of the PR and CD-based isolated models. Each table shows the five algorithms used, and the best model results are shown in bold. In the case of the DC-based models in Table 3, the best ML algorithm was ANN. However, all models except KNN obtained similar results (see Table 2 for details of the hyperparameters). Tables 4 and 5 show the models trained using the features extracted from the patient documents. From two used methods used for PR analysis, representation using the TF-IDF approach obtained the best result with an AUC of 60.8%, 9% lower than that presented by the CD-based models.

Table 6 shows the results of the type A data fusion models. As shown in Figure 1, each of these models receives both sources of information as input. The results are similar to those obtained in Table 3, which could be explained according to the information provided from CD data compared to patient documents since getting a higher performance with the first source is possible.

Table 2. Hyperparameters of ML Models.

Model	Main Hyperparameters
KNN	Algorithm: ball tree, leaf size: 5, number of neighbors: 2
LR	C:50, solver: lbfgs
SVM	C:500, kernel: polynomial
ANN	Layers: 1, Hidden neurons: 20, activation function: hyperbolic tangent
RF	Minimal number of splits: 2, number of estimators: 4

Table 3. Results for the CD Based Models.

Model	Sensitivity	Specificity	AUC
KNN	55.03.9%	64.87.9%	59.35.0%
LR	60.33.4%	70.95.1%	69.12.6%
SVM	67.26.4%	61.96.5%	67.33.7%
ANN	**62.72.8**%	**68.97.1**%	**69.92.3**%
RF	68.15.2%	57.69.1%	65.15.8%

Table 4. Results for the PRs based models using TF-IDF.

Model	Sensitivity	Specificity	AUC
KNN	65.93.6%	42.46.2%	54.83.5%
LR	**66.55.2**%	**47.97.8**%	**60.84.9**%
SVM	65.93.9%	41.96.8%	58.53.7%
ANN	62.54.5%	39.15.3%	51.03.5%
RF	80.33.4%	25.26.5%	53.25.6%

Table 5. Results for the MR Based Models using embeddings.

Model	Sensitivity	Specificity	AUC
KNN	60.23.9%	41.48.1%	50.84.4%
LR	63.24.8%	40.38.0%	52.35.6%
SVM	**73.63.6**%	**34.95.8**%	**56.04.1**%
ANN	57.06.0%	42.94.8%	50.33.5%
RF	70.64.7%	31.07.6%	50.15.9%

As mentioned in the experimental design, four algorithms were used for type B data fusion models (see Table 7), considering the results of Tables 3-5. From Table 3, the LR and SVM models were taken, which obtained better performance than the other models. ANN model was not considered because it was used in the architecture of Figure 3. Two models from Table 4 were taken into account because the TF-IDF method show better results than the embeddings one. Again, the algorithms taken were LR and SVM. There is no AUC metric for the voting classifier because it was not possible to estimate a probability function from the votes.

Table 6. Results for the data fusion models type I.

Model	Sensitivity	Specificity	AUC
KNN	60.05.1%	56.710.3%	57.46.0%
LR	**61.33.8**%	**68.45.8**%	**69.32.8**%
SVM	67.25.7%	62.16.4%	66.73.7%
ANN	64.53.3%	60.73.4%	66.92.3%
RF	78.62.7%	30.07.5%	59.32.6%

As mentioned in the experimental design for type II data fusion models, four algorithms were used considering the results of Tables 3-5. From Table 3, the LR and SVM models were taken, which obtained a better performance compared to other models. ANN model was not considered because it was used in the architecture of Figure 3. Only two models from Table 4 were taken from Tables 4 and 5 because the TF-IDF method showed better results than the embeddings method. Again, the algorithms taken were LR and SVM. For the voting classifier, and as in the previous scenario, there is no AUC metric because of the lacking possibility to estimate a probability function from the votes.

Table 7. Results for the data fusion models type II.

Model	Sensitivity	Specificity	AUC
Voting Classifier	64.24.0%	64.15.4%	–
Stacking Classifier	**71.34.3**%	**50.96.9**%	**67.23.6**%
ANN Fusion	71.83.5%	49.57.4%	65.63.9%

Comparing the performance of the four proposed schemes, the models based on CD were the ones that obtained the best results in terms of AUC value. Although it was also possible to classify patients by combining the information, the model's performance remained the same. However, a combination is an option to provide data to the diagnosis system support due to the lack of information unavailable due to different issues, such as the patients avoiding answering questions about HIV status, for example. In these cases, the interview of the patients and the PR created can be used with these tools to decide a patient's diagnosis.

Despite the result for sensitivity (73.6%) when PR was represented using embeddings, the AUC value was under the CD-based models at 69%. Comparable results reported an AUC of 71% and a sensitivity of 77% with a set of clinical variables similar to this study where an ANN obtained the best result [15]. In addition, in a similar scenario of limited infrastructure, reported sensitivity values of 97% [26]. Furthermore, specificities could have been better, according to 23% and 71% values, respectively. It is important to mention that, despite the similarities, the scenarios change with the involved variables and preprocessing stages.

Concerning the data fusion models, type A was better than B, with values of AUC of 69.3% and 67.2%, respectively. However, the sensitivity for the type B data fusion model reached 71.3%, comparable with more available tests based on smear microscopy. In addition, PR should have provided more information when the results were similar for CD-based models (AUC = 69.9%) and the best data fusion model (AUC = 69.3%), dropping by less than one percentage point. Nevertheless, the best PR-based proposal reached the best sensitivity value of the results with 73.6% (see Table 5). This aspect could be helpful when not all data are available. Health professionals can analyze different proposals for the specific diagnosis scenario and decide with more information than they hold with only CD or PR in an isolated mode.

The use of NLP in the TB diagnosis problem has been reported in terms of detection by employing cough signal analysis and embeddings in [50], differentiation of TB and Covid-19 from reported cases through text analysis in [51], and comparing a NLP-based image approach versus natural text generated by radiologists from chest radiographs (CXR) in [36]. However, works with the present approach based on text from the EMR has not been explored in the same manner as it was developed here. A similar work that employed the fusion of structured and unstructured data for Tuberculosis prediction with results of AUC with mean of 95.5% for different models was achieved in [52]. Nevertheless, this study was carried out with information from 692949 patients and more sophisticated clinical variables such as main complaint hemoptysis, cough, and test erythrocyte sedimentation rate, which is not comparable with the present work. Data fusion between text reports and images of CXR were used in a pre-trained model to establish a report from the available information, reaching a validation accuracy of 94%. A relevant difference can be found in terms of the number of images with around 3800 images from a known dataset [53]. Finally, exploratory works where the NLP technique has been used for extracting information on TB from EMR were reported in [**?**], but in those cases a classification for supporting the disease diagnosis was not implemented. With this, it is possible to see how the future and incremental work in this field can be improved.

Considering the specific context for the PTB diagnosis, where clinical signs and symptoms are considered as a first sight in the process, the possible patient is acknowledged as a suspect of having the disease. Then, according to the Colombian national protocol, it is necessary to include one of the three tests described in the introduction section. Smear microscopy is developed using sputum to visualize the mycobacterium in the microscopic test. However, due to the sufficient microorganisms that must be in the sample, the sensitivity reached by the test is between 40 and 60% [54]. Present results, with proposals based on embeddings and the type B data fusion, obtained sensitivity values of 73.6% and 71.3%, respectively, could be an option. In this case, the models can accomplish the activities of the health professionals when just smear microscopy is available. Another test with better sensitivity values, as culture in solid or liquid mediums, needs a time interval of two weeks and more sophisticated infrastructure [55], a scenario that is not always possible in low-income countries. Finally, with the best advantages regarding time and sensitivity results, the molecular test is not an option [56]. In Colombia, the national health system does not have more than twenty places with quality properties for this test. In addition, the solution is based on the transportation of the necessary samples, which can be compromised by inadequate shipping. Then, the infrastructure for this technology could be improved in developing countries such as Colombia. A tool based on the present proposal can provide information about the patient's condition to make a proper diagnosis and start treatment timely.

Limitations of the present study are related to the dataset size to propose the models. More samples are better for ML and, in general, AI applications that learn from data. However, the present approach was taken from a real scenario in the Hospital Santa Clara, which represents one of the institutions with the most PTB-treated patients. Unifying the data systems and transferring from different health centers with information previous to confirmed cases continues to be challenging in the same city. Availability of data with the same characteristics is difficult to establish, according to the national protocol for reporting TB-diagnosed patients. Moreover, it is important to mention that, with the advances and use of LLM architectures and its basic unit like the Transformer could be an option that is out of the scope of the present work, where preliminary information can be used. For this, a deeper analysis could be more specific, including the mentioned architectures. Something to pay attention to approaches like that is the lack of a considerable dataset to adjust the model, it will be necessary to increase the size of the data. This depends on the epidemiology aspects of the disease and medical center, which for this case is low but it is the most representative center in Bogota D.C. city.

Conclusion

The aim of this work is to generate alternatives for providing more information to health professionals based on computational intelligence tools. The proposal can be helpful in scenarios of making decisions for patients with suspected PTB. Analysis was implemented where limited resources to develop the diagnosis process are understood, and no commonly available diagnostic tests exist. This paper used two sources of information the employment of clinical data and the physicians' report found inside each patient's medical history. Different ML models were utilized with each source of data and two data fusion proposals based on text representation. The models that used only the PR obtained a performance of 9% lower than that obtained by the methods that used the CD but with better sensitivity results (73%). In the case of the data fusion models, their results could not overcome the use of clinical variables, and the sensitivity values were notable, too. However, as discussed, clinical variables are not always available, and certain considerations must be taken. So, the development of NLP techniques is significant as it can provide a great tool in places where there is no patient information or if the patient does not want to give personal information. The insights learned and algorithms used during the development of this work provide us with information for further studies to obtain a tool that can be used and integrated into the workflow of health personnel in the diagnosis of PTB.

Acknowledgments: Authors acknowledge the support of the Universidad del Rosario for funding this project. In addition, the contribution of the research incubator team Semillero en Inteligencia Artificial en Salud: Semill-IAS. In addition, the authors acknowledge the support of the Ministerio de Ciencia y Tecnología–Minciencias of Colombia, funded through project 123380762899 to develop this paper.

References

1. World Health Organization. Global Tuberculosis Report 2020. Technical Report Licence: CC BY-NC-SA 3.0 IGO, Geneva, 2020.
2. Group, T..G.S.; others. Tuberculosis and COVID-19 co-infection: description of the global cohort. *European Respiratory Journal* **2022**, *59*.
3. World Health Organization. Global Tuberculosis Report 2022. Technical Report Licence: CC BY-NC-SA 3.0 IGO, Geneva, 2022.
4. Teng, G.L.; Huang, Q.; Xu, L.; Chi, J.Y.; Wang, C.; Hu, H. Clinical features and risk factors of pulmonary tuberculosis complicated with pulmonary aspergillosis. *European Review for Medical and Pharmacological Sciences* **2022**, *26*, 2692–2701.
5. Shimoda, M.; Yoshiyama, T.; Okumura, M.; Tanaka, Y.; Morimoto, K.; Kokutou, H..; Osawa, T.; Furuuchi, K.; Fujiwara, K.; Ito, K.; Yoshimori, K.; Ohta, K. Analysis of risk factors for pulmonary tuberculosis with persistent severe inflammation: An observational study. *Medicine* **2022**, *101*, e29297.

6. Van't Hoog, A.; Viney, K.; Biermann, O.; Yang, B.; Leeflang, M.; Langendam, M. Symptom and chest radiography screening for active pulmonary tuberculosis in HIV negative adults and adults with unknown HIV status. *Cochrane Database of Systematic Reviews* **2022**.
7. Kwizera, R.; Katende, A.; Bongomin, F.; Nakiyingi, L.; Kirenga, B.J. Misdiagnosis of chronic pulmonary aspergillosis as pulmonary tuberculosis at a tertiary care center in Uganda: a case series. *Journal of Medical Case Reports* **2021**, *15*, 1–7.
8. Programa Nacional de Prevención y Control de la Tuberculosis, Ministerio de Salud. Informe de Evento Tuberculosis Año 2021. Technical report, Colombia, 2021.
9. Instituto Nacional de Salud. Boletin Epidemiológico Semanal, Semana 11. Technical report, Colombia, 2021.
10. Günther, G.; Ruswa, N.; Keller, P.M. Drug-resistant tuberculosis: advances in diagnosis and management. *Current Opinion in Pulmonary Medicine* **2022**, *28*, 211–217.
11. World Health Organization.; others. Rapid communication: key changes to the treatment of drug-resistant tuberculosis. Technical report, World Health Organization, 2022.
12. Ministerio de Salud y Protección Social. Resolución Número 0000227 de 2020. Technical report, Colombia, 2020.
13. Flores-Ibarra, A.A.; Ochoa-Vázquez, M.D.; Sánchez, T.G.A.. Estrategias diagnósticas aplicadas en la Clínica de Tuberculosis del Hospital General Centro Médico Nacional la Raza. *Rev Med Inst Mex Seguro Soc* **2016**, *54*, 122–127.
14. Ministerio de Salud y Protección Social. Plan Estratégico: Hacia el fin de la Tuberculosis, Colombia 2016-2025. Technical report, Colombia, 2016.
15. Orjuela-Cañón, A.D.; Jutinico, A.L.; Awad, C.; Vergara, E.; Palencia, A. Machine learning in the loop for tuberculosis diagnosis support. *Frontiers in Public Health* **2022**, *10*.
16. Sutton, R.T.; Pincock, D.; Baumgart, D.; Sadowski, D.; Fedorak, R.N.; Kroeker, K.I. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine* **2020**, *3*, 17.
17. Rajan, S.P.; Paranthaman, M. Artificial Intelligence in Healthcare: Algorithms and Decision Support Systems. *Smart Systems for Industrial Applications* **2022**, pp. 173–197.
18. Shortliffe, E.H.; Sepúlveda, M.J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **2018**, *320*, 2199–2200.
19. Jamshidi, M.B.; Daneshfar, F. A Hybrid Echo State Network for Hypercomplex Pattern Recognition, Classification, and Big Data Analysis. 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2022, pp. 007–012.
20. Amisha..; Malik, P.; Pathania, M.; Rathaur, V.K. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care* **2019**, *8*, 2328–2331.
21. Jamshidi, M.B.; Ebadpour, M.; Moghani, M.M. Cancer Digital Twins in Metaverse. 2022 20th International Conference on Mechatronics-Mechatronika (ME). IEEE, 2022, pp. 1–6.
22. Jamshidi, M.; Lalbakhsh, A.; Talla, J.; Peroutka, Z.; Hadjilooei, F.; Lalbakhsh, P.; Jamshidi, M.; La Spada, L.; Mirmozafari, M.; Dehghani, M.; others. Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *Ieee Access* **2020**, *8*, 109581–109595.
23. Jamshidi, M.B.; Talla, J.; Lalbakhsh, A.; Sharifi-Atashgah, M.S.; Sabet, A.; Peroutka, Z. A conceptual deep learning framework for COVID-19 drug discovery. 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2021, pp. 00030–00034.
24. Orjuela-Cañon, A.D.; Perdomo, O. Clustering proposal support for the COVID-19 making decision process in a data demanding scenario. *IEEE Latin America Transactions* **2021**, *19*, 1041–1049.
25. Dande, P.; Samant, P. Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review. *Tuberculosis (Edinb)* **2018**, *108*, 1–9.
26. Orjuela-Cañón, A.D.; Camargo Mendoza, J.E.; Awad García, C.E.; Vergara Vela, E.P. Tuberculosis diagnosis support analysis for precarious health information systems. *Computer Methods and Programs in Biomedicine* **2018**, *157*, 11–17.
27. Orjuela-Cañón, A.D.; de Seixas, J. Fuzzy-ART neural networks for triage in pleural tuberculosis. 2013 Pan American Health Care Exchanges (PAHCE), 2013, pp. 1–4.
28. Sung, S.F.; Chen, K.; Wu, D.P.; Hung, L.C.; Su, Y.H.; Hu, Y.H. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study. *International Journal of Medical Informatics* **2018**, *112*, 149–157.

29. Imler, T.D.; Morea, J.; Kahi, C.; Imperiale, T.F. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clin Gastroenterol Hepatol* **2013**, *11*, 689–694.
30. Wang, Y.; Luo, J.; Hao, S.; Xu, H.; Shin, A.Y.; Jin, B.; et al.. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. *Int J Med Inform* **2015**, *84*, 1039–1047.
31. Cai, T.; Giannopoulos, A.A.; Yu, S.; Kelil, T.; Ripley, B.; Kumamaru, K.K.; et al.. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics* **2016**, *36*, 176–191.
32. Doan, S.; Conway, M.; Phuong, T.M.; Ohno-Machado, L. Natural language processing in biomedicine: a unified system architecture overview. *Methods in Molecular Biology* **2014**, *1168*, 275–294.
33. Cui, H.; Fang, X.; Xu, R.; Kan, X.; Ho, J.C.; Yang, C. Multimodal fusion of ehr in structures and semantics: Integrating clinical records and notes with hypergraph and LLM. *arXiv preprint arXiv:2403.08818* **2024**.
34. Goel, A.; Gueta, A.; Gilon, O.; Liu, C.; Erell, S.; Nguyen, L.H.; Hao, X.; Jaber, B.; Reddy, S.; Kartha, R.; others. LLMs accelerate annotation for medical information extraction. Machine Learning for Health (ML4H). PMLR, 2023, pp. 82–100.
35. Abbasian, M.; Yang, Z.; Khatibi, E.; Zhang, P.; Nagesh, N.; Azimi, I.; Jain, R.; Rahmani, A.M. Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients. *arXiv preprint arXiv:2402.10153* **2024**.
36. Paul, H.Y.; Kim, T.K.; Lin, C.T. Comparison of radiologist versus natural language processing-based image annotations for deep learning system for tuberculosis screening on chest radiographs. *Clinical Imaging* **2022**, *87*, 34–37.
37. Zhao, F.; Zhang, C.; Geng, B. Deep Multimodal Data Fusion. *ACM Computing Surveys* **2024**.
38. Potdar, K.; Pardawala, T.S.; Pai, C.D. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications* **2017**, *175*, 7–9.
39. Romero Gómez, A.F.; Orjuela-Cañón, A.D.; Jutinico, A.L.; Awad, C.; Vergara, E.; Palencia, A. Preliminary Text Analysis from Medical Records for TB Diagnosis Support. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 2468–2471.
40. Zhu, W.; Zhang, W.; Li, G.Z.; He, C.; Zhang, L. A study of damp-heat syndrome classification using Word2vec and TF-IDF. 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2016, pp. 1415–1420.
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al.. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
42. Egger, R., Text Representations and Word Embeddings. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer International Publishing: Cham, 2022; pp. 335–361.
43. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. ICLR, 2013.
44. Uddin, S.; Haque, I.; Lu, H.; Moni, M.A.; Gide, E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports* **2022**, *12*, 6256.
45. Albahra, S.; Gorbett, T.; Robertson, S.; D'Aleo, G.; Kumar, S.V.S.; Ockunzzi, S.; Lallo, D.; Hu, B.; Rashidi, H.H. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. Seminars in Diagnostic Pathology. Elsevier, 2023, Vol. 40, pp. 71–87.
46. Ozer, M.E.; Sarica, P.O.; Arga, K.Y. New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. *Omics: a journal of integrative biology* **2020**, *24*, 241–246.
47. Raschka, S. *Python Machine Learning: Unlock Deeper Insights Into Machine Learning with this Vital Guide to Cutting-edge Predictive Analytics*; Community experience distilled, Packt Publishing, 2015.
48. Chollet, F.; others. Keras. https://github.com/fchollet/keras, 2015.
49. Awan, S.E.; Bennamoun, M.; Sohel, F.; Sanfilippo, F.M.; Dwivedi, G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC heart failure* **2019**, *6*, 428–435.
50. Pahar, M.; Theron, G.; Niesler, T. Automatic Tuberculosis detection in cough patterns using NLP-style cough embeddings. 2022 International Conference on Engineering and Emerging Technologies (ICEET). IEEE, 2022, pp. 1–6.
51. Pholo, M.D.; Hamam, Y.; Khalaf, A.B.; Du, C. Differentiating Between COVID-19 and Tuberculosis Using Machine Learning and Natural Language Processing. *Revue d'Intelligence Artificielle* **2022**, *36*.

52. Wang, M.; Lee, C.; Wei, Z.; Ji, H.; Yang, Y.; Yang, C. Clinical assistant decision-making model of tuberculosis based on electronic health records. *BioData Mining* **2023**, *16*, 11.
53. Naz, I.; Iftikhar, S.; Zahra, A.; Zainab, S. Report Generation of Lungs Diseases from Chest X-Ray Using NLP. *International Journal of Innovations in Science and Technology* **2022**, *3*, 223–33.
54. Lewinsohn, D.M.; Leonard, M.K.; LoBue, P.A.; Cohn, D.L.; Daley, C.L.; Desmond, E.; Keane, J.; Lewinsohn, D.A.; Loeffler, A.M.; Mazurek, G.H.; others. Official American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention clinical practice guidelines: diagnosis of tuberculosis in adults and children. *Clinical Infectious Diseases* **2017**, *64*, e1–e33.
55. Ghazvini, K.; Yousefi, M.; Firoozeh, F.; Mansouri, S. Predictors of tuberculosis: Application of a logistic regression model. *Gene Reports* **2019**, *17*, 100527.
56. Berra, T.Z.; Gomes, D.; Ramos, A.C.V.; Alves, Y.M.; Bruce, A.T.I.; Arroyo, L.H.; Santos, F.L.d.; Souza, L.L.L.; Crispim, J.d.A.; Arcêncio, R.A. Effectiveness and trend forecasting of tuberculosis diagnosis after the introduction of GeneXpert in a city in south-eastern Brazil. *Plos one* **2021**, *16*, e0252375.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.