applied sciences

MDPI

Article

Data Fusion of Medical Records and Clinical Data to Enhance Tuberculosis Diagnosis in Resource-Limited Settings

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

[1] School of Medicine and Health Sciences, Universidad del Rosario, Bogota 111221, Colombia
[2] Fundación Santa Fe de Bogotá, Bogota 110111, Colombia; andres.romero@fsfb.org.co
[3] Biomedical Engineering, Universidad Antonio Nariño, Bogota 110311, Colombia; ajutinico@uan.edu.co
[4] Subred Integrada de Servicios de Salud Centro Oriente, Bogota 111711, Colombia; carlosawad@gmail.com (C.E.A.); angelicapalenciab@gmail.com (M.A.P.)
[5] Hospital Universitario Nacional, Bogota 111321, Colombia; 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 that has been declared a global emergency by the World Health Organization and remains one of the top ten causes of death worldwide. TB diagnosis is particularly challenging in developing countries, where limited infrastructure for detection and treatment complicates efforts to control the disease. These resource constraints are especially critical in remote areas with few mechanisms for timely diagnosis, which is essential for effective patient management. Artificial intelligence (AI) has emerged as a valuable tool in supporting health professionals by enhancing diagnostic processes. This paper explores the use of natural language processing (NLP) techniques and machine learning (ML) models to facilitate TB diagnosis in settings where robust data infrastructure is unavailable. Two distinct data sources were analyzed: text extracted from electronic medical records (EMRs) and patient clinical data (CD). Four different ML-based approaches were implemented: two models using each data source independently and two data fusion models combining both sources. The relevance of these strategies was assessed in collaboration with physicians to ensure their practical applicability in clinical decision-making. The results of the data fusion models were compared to determine which source provided more valuable diagnostic information. The best-performing model, which relied solely on CD, achieved a sensitivity of 73%, outperforming smear microscopy, which typically ranges from 40% to 60%. These findings underscore the importance of analyzing physicians' reports and assessing the availability of such information alongside structured clinical data. This approach is particularly beneficial in resource-limited settings, where access to comprehensive clinical data may be restricted.

Keywords: artificial intelligence; tuberculosis diagnosis; data fusion

Academic Editor: Anca Udristoiu

Received: 2 March 2025
Revised: 30 March 2025
Accepted: 30 March 2025
Published: 13 May 2025

Citation: Orjuela-Cañón, A.D.; Romero-Gómez, A.F.; Jutinico, A.L.; Awad, C.E.; Vergara, E.; Palencia, M.A. Data Fusion of Medical Records and Clinical Data to Enhance Tuberculosis Diagnosis in Resource-Limited Settings. *Appl. Sci.* **2025**, *15*, 5423. https://doi.org/10.3390/app15105423

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 one of the top ten causes of death worldwide. Moreover, in 2023, it returned as the leading cause of death from a single infectious agent, overshadowed by COVID-19 in the 2020 to 2023 period [1]. In developing countries, the situation is worse due to the risk of developing TB after the infection is

associated with sociopolitical issues and healthcare infrastructure, representing a significant public health challenge. 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 4.6% between 2020 and 2023, due to the impact of the COVID-19 pandemic. In addition, it was estimated that more than ten million people fall ill with TB, marking an increase in this number since 2021 [1,2].

It is estimated that one in four people worldwide is infected with latent TB, with the potential to develop active TB, especially in individuals with compromised immune systems—such as those with HIV, diabetes, malnutrition, tobacco use, or homelessness, who have a 5–10% chance [1,3,4]. TB can affect any organ in the body but primarily targets the lungs, a condition known as pulmonary TB (PTB), which presents symptoms such as a severe cough lasting more than three weeks, chest pain, and coughing up blood or sputum [5,6].

This study analyzed a case using data from a developing country, specifically Colombia. In 2023, Colombia reported a tuberculosis (TB) incidence rate of 32.64 per 100,000 population, which increased to 40.85 in 2024—marking a 13% and 25% rise in reported pulmonary TB (PTB) cases compared to the previous years [1,7]. This growing burden highlights the persistent disparities in resource distribution within the public health sector. Addressing these challenges requires the development of alternative strategies to enhance disease management, facilitate early detection, and expedite the initiation of anti-PTB treatment.

National protocols allow healthcare professionals to initiate treatment based on clinical evaluation [8], even in the absence of bacteriological confirmation, to prevent disease transmission and progression in patients based on three primary diagnostic tests: smear microscopy, molecular tests, and culture. Smear microscopy is the simplest and most affordable test, providing results within a short time. However, its sensitivity is low, ranging between 40% and 60%, depending on the quality of the sample. In contrast, molecular tests offer higher sensitivity, exceeding 90%, and can deliver results within hours. The main drawback of this method is the need for specialized equipment and trained professionals, making it more costly than smear microscopy. Finally, culture is the most reliable diagnostic method, offering high sensitivity and specificity. However, it requires skilled personnel, expensive infrastructure, and a processing time of at least two to three weeks [8,9]. Each test has its advantages and limitations, and depending on resource availability, a patient may undergo one, two, or all three tests. Nonetheless, the time and costs associated with these diagnostic methods create accessibility barriers in certain regions. Therefore, developing new low-cost, rapid technologies is essential to support healthcare professionals in diagnosing the disease more efficiently [10].

Despite the availability of current technologies and diagnostic protocols for PTB, some regions in the country still lack the necessary resources to implement these procedures effectively. In many cases, healthcare professionals rely on traditional tools, while access to laboratories and advanced diagnostic equipment remains limited and often delayed. Previous research conducted by the same team has explored alternative approaches for addressing these challenges in [11–13]. In this study, an extension was based on a previously reported method, where medical staff documented patient consultations with detailed descriptions of medical findings and terminology. This documentation, commonly used in the traditional Colombian healthcare system, can serve as an additional data source to support the diagnostic process.

In recent years, artificial intelligence (AI) has been increasingly utilized in medicine to support decision-making [14,15]. AI-powered systems provide healthcare professionals with valuable insights, enabling them to enhance diagnostic accuracy and efficiency. These

systems can process large amounts of data and apply machine learning (ML) techniques to specific medical tasks [16,17]. One of the key advantages of AI-based tools is the cost-effectiveness and accessibility for frontline healthcare workers, making them particularly useful in situations where conventional diagnostic methods are unavailable [18]. AI has been successfully applied in various medical fields, including cancer detection [19], COVID-19 diagnosis and treatment, and drug discovery through deep neural networks [20–23]. In the case of TB, studies [12,24] have demonstrated how artificial neural networks (ANNs) can be trained to diagnose TB using clinical data (CD). Additionally, research in [22,25] explores ANN-based clustering techniques to categorize populations into three risk groups (high, medium, and low risk) showing promising results in TB and COVID-19 risk assessment.

Natural language processing (NLP) is a branch of AI that enables the analysis of unstructured text and can be used to develop decision-support tools. For example, NLP has been applied to build AI systems that assist with tasks such as retrieving relevant information [26], determining patient eligibility and tracking [27,28], and diagnosing diseases [29]. These systems typically rely on data from electronic medical records (EMRs), including clinical notes (CNs), laboratory results, and imaging data. Research has shown that the most effective NLP models are those that learn patterns from data using ML techniques [30]. Since 2021, the rise of large language models (LLMs) has driven advancements in generative models and new approaches to text processing. However, these methods face challenges related to the specific dataset used for model development and the manual annotations provided by expert medical professionals [31,32]. Examples of NLP applications in healthcare include assessing dietary risks for diabetes patients [33] and comparing NLP-generated labels based on medical imaging with expert TB screening in chest X-rays [34]. Additionally, NLP and LLM have been used for data fusion in medical contexts [31,32,35].

To enhance PTB diagnosis, we proposed four predictive modeling schemes utilizing two data sources. The first source comprises clinical data (CD) from 151 patients, including variables such as HIV status, geographic location within the city, and sex. Additionally, clinical notes (CNs) from physicians' reports (PRs) during patient consultations as part of routine practice were included. Two schemes integrate both CD and CN, while the other two process them separately. Each scheme employs five machine learning (ML) algorithms: support vector machine (SVM), K-nearest neighbor (KNN), logistic regression (LR), random forest (RF), and artificial neural networks (ANNs). Since CN consists of unstructured text, NLP techniques were applied to convert these reports into numerical representations, allowing ML algorithms to extract meaningful patterns. This approach addresses the challenges faced by resource-limited settings, where access to advanced diagnostic infrastructure is scarce. By leveraging data fusion techniques and computational intelligence, the proposed models can support TB diagnosis in regions lacking well-developed healthcare systems, improving early detection and intervention.

2. Materials and Methods

Figure 1 presents an overview of the implemented methodology. First, data were collected from Hospital Santa Clara in Bogota, Colombia. Next, a preprocessing and cleaning phase was conducted to prepare the data before extracting the features used as input for the algorithms. For this, four different schemes were designed, each employing five ML models. To assess the model performance, the stratified k-fold cross-validation technique was established. The following sections provide a detailed explanation of each step illustrated in Figure 1.

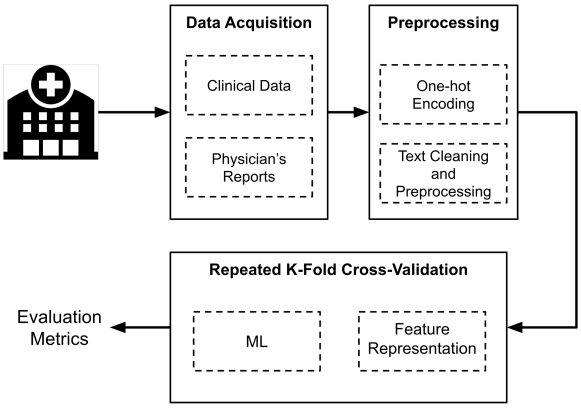

Figure 1. General scheme of the methodology implemented.

2.1. Dataset

The data were collected between 2017 and 2019 at Hospital Santa Clara in Bogotá D.C., Colombia, which is part of the Middle-East Integrated Health Services Subnetwork (translated from Spanish: *Subred Integrada de Servicios de Salud Centro Oriente*). The study was approved by the subnetwork's ethics committee under Act No 316 of 24 May 2021. Approval was granted based on the anonymization of data, which included only population-related variables and posed no risk to subjects. Informed consent was not required, as all data were retrospective and fully anonymized.

Clinical data (CD) variables were extracted from records of the institutional PTB program, while clinical notes (CNs) from the physician reports (PRs) were obtained from electronic medical records, which are standardized according to the hospital's record acquisition system. In Colombia, EMRs remain a challenge due to limitations in basic guidelines, policies defining required data fields, and issues related to interoperability [36].

The dataset included clinically suspected PTB cases, with microbiological tests performed to confirm the disease. These tests identified 116 confirmed PTB cases and 35 non-PTB patients from an initial dataset of 233 individual records. The remaining cases were excluded due to missing PTB confirmation or unavailable PR data.

For PR extraction, when applicable, dates related to diagnostic tests and the initiation of treatment were prioritized. The temporal placement of PR was considered crucial, as it contained information indicative of PTB. Therefore, up to five PRs recorded within 30 days prior to these key dates were extracted, analyzed, and summarized. This was necessary because not all EMR data were directly related to PTB but rather encompassed the patient's entire medical history.

Regarding CD, data were collected by physicians involved in the institutional PTB program, as well as members of the research team. The extracted variables included (i) geographic location within Bogotá, (ii) sex, (iii) human immunodeficiency (HIV) status, (iv) antiretroviral treatment (ART) status, (v) population risk factors, such as homelessness, migrant status, displacement, or indigenous background. These variables were selected based on recommendations from physicians and were encoded using binary representation [37]. Table 1 provides an overview of the variables and their possible values.

Table 1. Variables used for the CD source data.

Variable	Values
Sex	Male
	Female

Table 1. *Cont.*

Variable	Values
Population	Homeless
Risk	Native
Factors	Exile
	Immigrant
	Prison
	Victim of Violence
	Other
Geographical	Antonio Nariño
City	Barrios Unidos
Location	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	Yes
Status	No
	Unknown
Antiretroviral	Yes
Treatment	No
Status	Unknown

2.2. Preprocessing

During the preprocessing and data cleaning phase, an exploratory analysis was conducted to identify and correct any errors in data acquisition. Based on this analysis, a final dataset of 151 patients was constructed, as some cases had to be excluded due to missing EMRs or discrepancies between recorded diagnostic test dates and those found in the EMR system.

Additionally, Bogotá's population was recategorized into five zones (north, south, east, west, and areas outside the city) based on geographical patients' places of residence. This zoning ensured a significant number of patients per region. The assigned localities within each zone correspond to those designated by the city for the four integrated health services subnetworks, allowing for the inclusion of indirect sociodemographic information.

For NLP preprocessing, the PR involved text cleaning through a three-step process. First, all words were converted to lowercase. Second, stopwords were removed using a standard stopword list to filter out prepositions, adverbs, articles, and conjunctions. Finally, a customized word removal process was applied to eliminate repetitive terms frequently found in the TB diagnosis process, such as medication units, institution-specific terms related to patients, and other TB-related words requiring lemmatization and stemming. The selection of these additional words was conducted in collaboration with healthcare professionals to ensure relevance.

Additionally, clinical reports often contain terms related to information or events unrelated to PTB [38]. However, the impact of these terms was minimized through post-processing. Non-alphabetic characters, including accents, punctuation marks, numbers, and other symbols, were also removed. Furthermore, double spaces and line breaks were eliminated to refine the text and prevent unnecessary elements from affecting the algorithms. More details on dataset construction from EMR and a preliminary analysis of text content can be found in [38]. For practical issues, the term *patient document* refers to the combined and preprocessed text from the five PR associated with each patient.

2.3. Feature Representation

The CD source data were structured using nominal variables represented as numerical vectors. To achieve this, one-hot encoding was applied, assigning unique binary values to each category of the original variable. For example, the HIV status variable has three possible values: positive, negative, and unknown. Consequently, three separate columns were created, each representing one of these values. A binary value of one (1) was assigned when the condition was present, and zero (0) otherwise. This numerical representation enables ML algorithms to be trained using categorical variables from the CD without implying any specific order or preference among them.

For NLP applications, computers process language in the form of text extracted from patients' documents. However, these texts need to be converted into numerical representations for ML algorithms to analyze and learn from the data. To achieve this, two different text representation methods were explored, where each document is transformed into a vector that captures its content.

Term Frequency-Inverse Document Frequency: The first method employs term frequency-inverse document frequency (TF-IDF) [39]. This metric helps emphasize words that frequently appear in a document while penalizing those that occur across multiple documents. The informativeness of a word is inversely related to the number of documents in which it appears, as words that are common across documents contribute less to classification. Equation (1) shows how $IDF(t)$ 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 calculating the IDF for all terms, the TF-IDF is obtained by multiplying the term frequency (TF) of each item in a document by its corresponding IDF value. As a result, each document is represented by a vector containing the TF-IDF values for all terms in the dataset. Additionally, L2 normalization was applied to each vector to ensure that the sum of the squares of its elements equals 1.

Moreover, TF-IDF terms are not limited to single words; n-grams can also be used. N-grams are contiguous sequences of elements—for example, a 2-g consists of two consecutive

words in a document. This approach captures language structure and enhances contextual understanding. Several n-gram combinations were explored to construct the vocabulary. Different combinations of *n-grams* were explored to build the vocabulary: (i) 1-g (a single term), (ii) 1-g and 2-g combination, (iii) 2-g, (iv) 1-g, 2-g, and 3-g combination, (v) 3-g. Since the vocabulary can become extensive, the vector size was constrained within a specified $DF(t)$ range, with the minimum and maximum values optimized during the process. Additionally, the vocabulary was capped at 1000 terms. However, given the high dimensionality relative to the available samples, dimensionality reduction was applied using truncated singular value decomposition (SVD), with the number of components also optimized.

Version 1.0 of the *Scikit Learn* library was used to compute the TF-IDF vector for each document [40]. The vectorization process followed the parameters described earlier, including word-based analysis, *n-grams* ranging unigrams to trigrams, and a maximum of 1000 features. As mentioned in the previous subsection, stop words were excluded during preprocessing based on the Spanish language.

Embeddings Representation: The second method employed embeddings representations. In this approach, each term is represented by a vector, where words with similar meanings or those used in similar contexts have closely related vector representations. There are different ways to generate embeddings: they can be learned from an ML model during training on a specific task or obtained in an unsupervised manner from document statistics [41]. In this work, the Word2Vec model was used to generate embeddings [42]. Two algorithms were applied: continuous bag of words (COBW), which predicts a target word based on surrounding words within a window, and skip-gram, the reverse task by predicting context words for a given target word using unsupervised learning techniques. Both methods rely on neural networks, including layers and weights, to predict words and generate meaningful word embeddings.

Finally, while each term in a document is represented as a vector, a single vector representation for the entire document is required to train traditional ML algorithms. To achieve this, a pooling step was applied using two different operations: maximum pooling, which selects the maximum value for each dimension across all word vectors in the document, and mean pooling, which computes the average of the word vectors. As a result, each document was represented by a vector containing either the maximum or the mean of the embeddings corresponding to its words.

2.4. Machine Learning Models

Before discussing the experimental design, it was necessary to briefly explain the ML algorithms used. As mentioned above, the five algorithms employed were KNN, LR, SVM, ANN, and RF. These models were trained for binary classification to distinguish between patients with confirmed TB and those without TB.

KNN is an instance-based learning classifier that does not attempt to build a general internal model. Instead, it stores instances of the training data and classifies new samples based on the k (a specified number) nearest neighbors. Classification is determined either by a majority vote among the nearest neighbors or by assigning weights inversely proportional to the distance from the point being classified [43].

The regression task developed by the LR model consists of optimizing a cost function. It models the likelihood of a binary outcome (in this case, TB detection) using a logistic function [40].

For ANN, different architectures can be applied, In this study, a single hidden layer with a set of neurons was used. Each neuron performs a weighted sum of its inputs and

applies an activation function, which adds nonlinearity to its output. The model learns the optimal weights from the training set and generalizes them to new inputs [44].

Another approach for classification can be implemented through the use of SVM. This model aims to maximize the margin, defined as the distance between the separation hyperplane and the closest labeled training samples, known as support vectors [45]. These support vectors contribute to the distinguishing between the TB-detected and non-detected classes.

The hyperparameter optimization was performed using a grid search, systematically evaluating different parameter combinations to identify the best configuration based on training performance for each model. For the SVM, the parameters tested included the regularization parameter C, kernel type, and gamma. The KNN model was optimized by varying the number of neighbors, weight function, algorithm type, and leaf size. In the case of the LR model, the regularization parameter C and optimizer were adjusted. For the ANN, the optimization process involved tuning the number of neurons and layers, batch size, number of epochs, activation function type, and learning rate. Finally, for the RF model, the number of estimators and the sample-splitting strategy were explored.

Finally, the RF model is an ensemble training method that combines multiple decision trees, each trained independently. While a single tree may not be sufficient for accurate classification, RF improves performance by aggregating multiple weak classifiers to enhance prediction accuracy [46]. All models were implemented using the *Scikit Learn* library, except for ANN, which was implemented using the Keras library and its version 2.7.0rc0 [47]. For each case, the hyperparameters were optimized through a grid search strategy, where different values were tested, and the best-performing configurations were selected.

2.5. Experimental Design

As mentioned earlier, five ML schemes were proposed for classification, as illustrated in Figure 2. These schemes include two approaches that use PR and CD separately and two fusion schemes that combine both sources of information: Type A and Type B data fusions. The ML approaches were applied to the PR-based, CD-based, and Type A fusion models. In contrast, the Type B fusion models were constructed based on the performance of the PR-based and CD-based approaches. This allows for selecting the best-performing ML algorithms for the classification task.

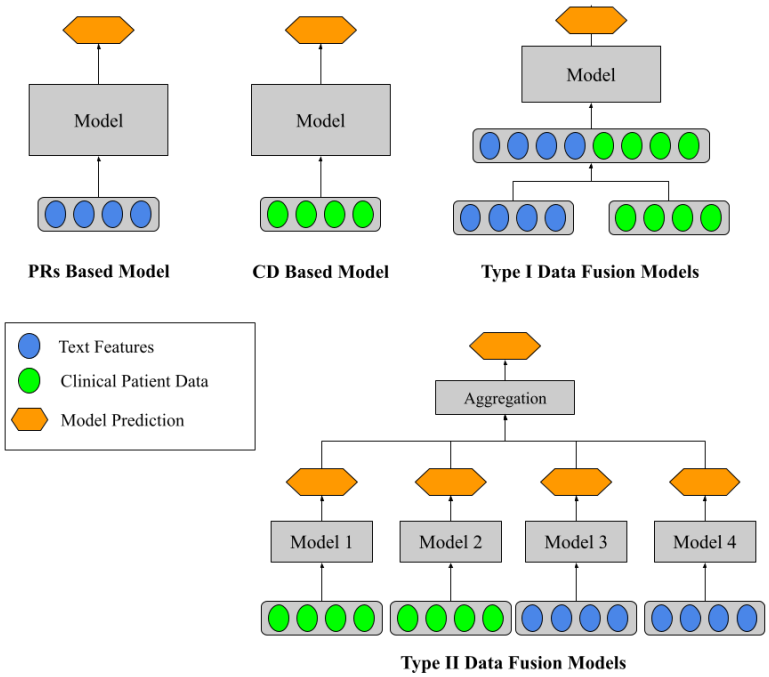

Figure 2. The four proposed ML classification schemes.

The Type B data fusion model incorporates an aggregation layer with two different strategies: (a) majority voting, where the final classification is determined by a simple majority vote among the four classifiers. In case of a tie, priority is assigned in descending order from model 1 (highest priority) to model 4 (lowest priority); (b) stacking classifier, where an ML algorithm is used as the aggregation layer. LR was chosen for this task, where the inputs are the probability outputs or decision functions (depending on the type of ML model) from the four classifiers.

Additionally, a different model was introduced alongside the previously mentioned approaches. This model consists of a neural network architecture, as illustrated in Figure 3. The design incorporates two separate networks, each dedicated to learning from PR and CD data sources independently during the training process. These networks are then concatenated, allowing a subsequent layer to handle information fusion. Finally, the output layer consists of a single neuron with a sigmoid activation function, which performs the binary classification task.

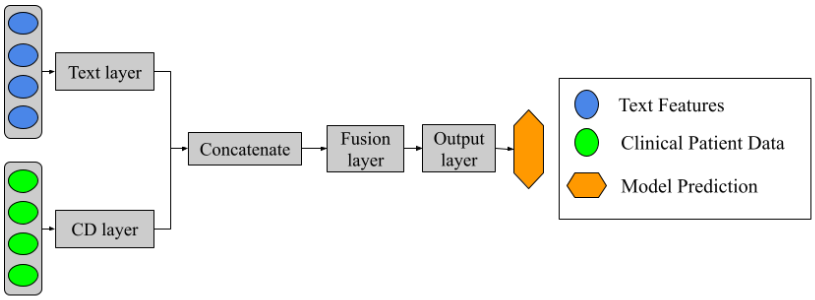

Figure 3. ANN used for information fusion (ANN fusion).

Stratified *k*-fold cross-validation (SKCV) was used to select the best parameters for text representation (TF-IDF and embeddings) and to optimize the hyperparameters of the ML models, as illustrated in Figure 1. In SKCV, the dataset was split into multiple partitions for training and validation, ensuring that each set remained distinct. The term "stratified" indicates that the class distribution was preserved across all partitions. Given the database size, a three-fold approach was chosen to maintain a sufficient number of samples in the validation set for an accurate model evaluation. To ensure statistical reliability, SKCV was repeated ten times, and the mean and standard deviation of the performance metrics across all repetitions were used to assess and compare the models.

The parameters explored for obtaining the TF-IDF measure included different combinations of *n-grams* as well as the maximum and minimum values of $DF(t)$ required for a term to be included in the vocabulary. Additionally, the number of components used for dimensionality reduction was examined alongside the computed TF-IDF values. Both TF-IDF calculation and dimensionality reduction were fitted using the training set before being applied to both the training and testing sets. This approach ensures that when predicting the diagnosis of a document, the TF-IDF method does not incorporate vocabulary from the test set, preserving the integrity of the model evaluation. By doing so, test documents accurately simulate the addition of new, unseen inputs.

For the Word2Vec models training, the COBW and skip-grams neural network models were implemented using the training set. Several parameters were explored, including the window size and the minimum frequency a word needed to appear to be included in the vocabulary. Additionally, during the pooling step, both maximum pooling and mean pooling operations were applied. For both feature extraction methods, each document was represented by a vector, followed by a normalization step, where values were transformed into a range between -1 and 1 to ensure consistency across representations.

Furthermore, the hyperparameters of the ML algorithms were explored using a random grid search in a heuristic manner until the best results were achieved. For models with a small set of hyperparameters, every possible option was systematically tested whenever feasible, such as LR, ANN, and SVM kernel models. For the fusion models, the same methodology was applied; however, the exploration process was shorter, as it was guided by the results obtained from the PR and CD-based models. Additionally, it is important to note that the text representation parameters and the ML hyperparameters were optimized together, ensuring an effective combination for classification performance.

The metrics used to evaluate the results of SKCV were sensitivity, specificity, and area under the curve (AUC), as they are particularly relevant in healthcare systems [48]. These metrics were computed for each testing set and then averaged. To ensure robust evaluation, the SKCV process was repeated ten times, and the mean and standard deviation of the three metrics were calculated across all repetitions, providing insight into the variability of the models. Additionally, given the imbalanced dataset, the training was weighted according to the number of samples in each class, ensuring that the metric accurately reflects the dataset distribution. Finally, the same training and testing sets were consistently used across all models to enable fair comparisons.

3. Results and Discussion

Tables 2–4 present the results of the PR and CD-based isolated models. Each table lists the five ML algorithms used, with the best-performing model highlighted in bold. For the CD-based models (Table 2), ANN achieved the best performance. However, all models—except KNN—produced similar results (see Table 5 for hyperparameter details). Tables 3 and 4 display the models trained using features extracted from patient documents. Among the two PR analysis methods, the TF-IDF representation yielded the best result, with an AUC of 60.8%, a 9% lower than that of the CD-based models. Figures 4 and 5 exhibit the dispersion of the results through boxplot graphics.

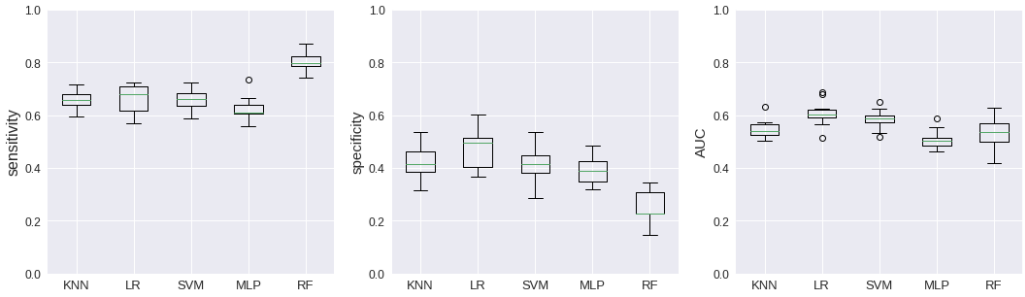

Figure 4. Results for the PR-based models using TF-IDF.

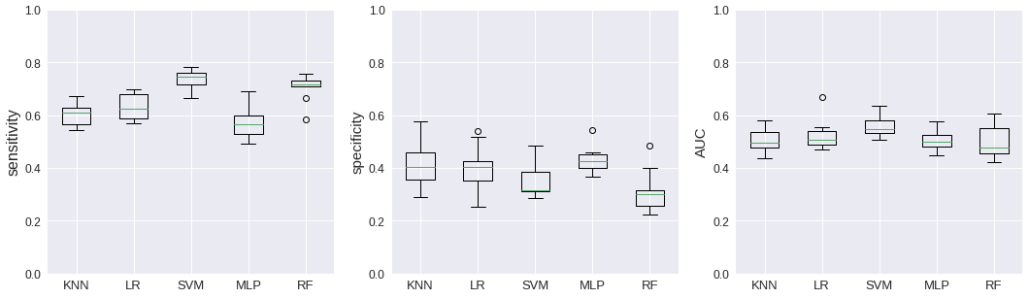

Figure 5. Results for the PR-based models using embeddings.

Table 6 presents the results of the Type A data fusion models. As shown in Figure 2, these models incorporate both sources of information as input. Figure 6 visualizes the ROC

curve for the best performance of the models in the cross-validation technique. The results are comparable to those in Table 2, likely due to the greater predictive value of CD data compared to patient documents, which may contribute to the higher performance observed with CD-based models.

Table 2. Results for the CD-based models.

Model	Sensitivity	Specificity	AUC
KNN	55.0 ± 3.9%	64.8 ± 7.9%	59.3 ± 5.0%
LR	60.3 ± 3.4%	70.9 ± 5.1%	69.1 ± 2.6%
SVM	67.2 ± 6.4%	61.9 ± 6.5%	67.3 ± 3.7%
ANN	**62.7 ± 2.8%**	**68.9 ± 7.1%**	**69.9 ± 2.3%**
RF	68.1 ± 5.2%	57.6 ± 9.1%	65.1 ± 5.8%

Table 3. Results for the PR-based models using TF-IDF.

Model	Sensitivity	Specificity	AUC
KNN	65.9 ± 3.6%	42.4 ± 6.2%	54.8 ± 3.5%
LR	**66.5 ± 5.2%**	**47.9 ± 7.8%**	**60.8 ± 4.9%**
SVM	65.9 ± 3.9%	41.9 ± 6.8%	58.5 ± 3.7%
ANN	62.5 ± 4.5%	39.1 ± 5.3%	51.0 ± 3.5%
RF	80.3 ± 3.4%	25.2 ± 6.5%	53.2 ± 5.6%

Table 4. Results for the PR-based models using embeddings.

Model	Sensitivity	Specificity	AUC
KNN	60.2 ± 3.9%	41.4 ± 8.1%	50.8 ± 4.4%
LR	63.2 ± 4.8%	40.3 ± 8.0%	52.3 ± 5.6%
SVM	**73.6 ± 3.6%**	**34.9 ± 5.8%**	**56.0 ± 4.1%**
ANN	57.0 ± 6.0%	42.9 ± 4.8%	50.3 ± 3.5%
RF	70.6 ± 4.7%	31.0 ± 7.6%	50.1 ± 5.9%

Table 5. 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 6. Results for the data fusion models Type A.

Model	Sensitivity	Specificity	AUC
KNN	60.0 ± 5.1%	56.7 ± 10.3%	57.4 ± 6.0%
LR	**61.3 ± 3.8%**	**68.4 ± 5.8%**	**69.3 ± 2.8%**
SVM	67.2 ± 5.7%	62.1 ± 6.4%	66.7 ± 3.7%
ANN	64.5 ± 3.3%	60.7 ± 3.4%	66.9 ± 2.3%
RF	78.6 ± 2.7%	30.0 ± 7.5%	59.3 ± 2.6%

As outlined in the experimental design, four algorithms were selected for the Type B data fusion models (see Table 7), based on the results from Tables 2–4. From Table 2, LR and

SVM were chosen, as they demonstrated better performance compared to the other models. The ANN model was excluded since it was already incorporated into the architecture shown in Figure 3. Additionally, two models from Table 3 were selected, as the TF-IDF method outperformed the embeddings approach. Again, LR and SVM were chosen as the most effective algorithms. It is important to note that the AUC metric is not available for the voting classifier, as it was not possible to estimate a probability function based on the votes. Figure 7 visualizes the ROC curve for the best performance of the models in the cross-validation technique.

Table 7. Results for the data fusion models Type B.

Model	Sensitivity	Specificity	AUC
Voting Classifier	64.2 ± 4.0%	64.1 ± 5.4%	64.1 ± 2.5%
Stacking Classifier	**71.3 ± 4.3**%	**50.9 ± 6.9**%	**67.2 ± 3.6**%
ANN Fusion	71.8 ± 3.5%	49.5 ± 7.4%	65.6 ± 3.9%

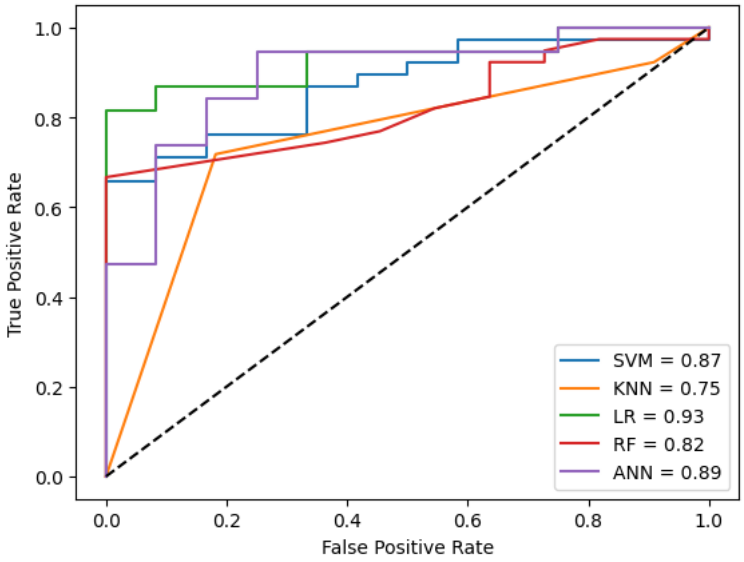

Figure 6. ROC curve for the fusion Type A.

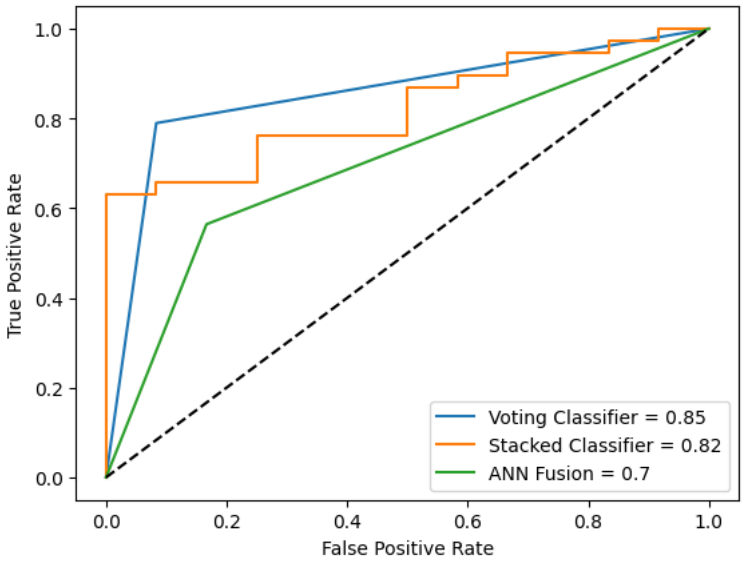

Figure 7. ROC curve for the fusion Type B.

When comparing the performance of the four proposed schemes, the CD-based models achieved the highest AUC values, indicating the best results. While combining information from different sources also enabled patient classification, the overall model performance remained unchanged. However, data fusion remains a valuable approach for supporting diagnostic decision-making, particularly in cases where certain patient information is unavailable. For example, some patients may avoid disclosing their HIV status or other critical details. In such situations, patient interviews and PR data can be leveraged to enhance diagnostic accuracy and compensate for missing information.

Despite achieving a sensitivity of 73.6% when PR was represented using embeddings and SVM model, and a value of 80.3% for TF-IDF representation and RF model, the AUC values (56.0% and 53.2%) remained lower than that of the CD-based models (see Table 2). Comparable studies have reported an AUC of 71% and a sensitivity of 77% using a similar set of clinical variables, where an ANN model achieved the best performance [11]. Additionally, in a resource-limited setting, sensitivity values as high as 97% have been reported [12]. However, specificity values in these studies were relatively low, recorded at 23% and 71%, respectively. It is important to note that, despite these similarities, differences in the variables used and preprocessing techniques influence the final results.

Regarding the data fusion models, Type A outperformed Type B, achieving an AUC of 69.3% compared to 67.2%. However, the sensitivity of the Type B model reached 71.3%, making it comparable to widely used smear microscopy tests, which typically have a sensitivity ranging from 40% to 60%, the first feasible diagnostic option in the Colombian protocol. Ideally, PR should have provided additional information, yet the results remained similar between CD-based models (AUC = 69.9%) and the best data fusion model (AUC = 69.3%), with a drop of less than one percentage point. Notably, the best PR-based approach achieved the highest sensitivity (73.6%) (see Table 4), which is particularly valuable when not all data sources are available. In such cases, health professionals can assess multiple diagnostic models and make more informed decisions rather than relying solely on CD or PR in isolation.

A distinctive aspect of this research is its emphasis on leveraging available data to train models in resource-constrained healthcare settings. Many developing countries face significant challenges, including inconsistent health data collection, restricted data availability, and shortages of healthcare personnel, all of which impact the development of reliable models. This research is part of a larger initiative aimed at developing computational intelligence tools to assist healthcare professionals in diagnosing and managing suspected PTB cases more effectively.

The tuberculosis diagnosis program at Hospital Santa Clara relies on two distinct sources of data. The first is structured data collected by nursing professionals, who record acquisition variables based on test results and patient interviews. The second source consists of unstructured data documented by physicians in text reports during patient appointments. These two sources are largely independent and collected at different moments. While they are eventually integrated into the health record through the information system, timely TB diagnosis is crucial for isolating suspected patients and initiating anti-TB therapy.

This study aims to determine whether both data sources are complementary or if one alone is sufficient for diagnosis. Additionally, the proposal seeks to provide an alternative for remote areas where limited digital health infrastructure makes maintaining a comprehensive electronic health record challenging. To enhance the paper, these details have been highlighted in red in the discussion section, following the previous comment.

A key consideration in processing data for NLP applications is the comparison between TF-IDF and embedding techniques. In this study, TF-IDF demonstrated superior performance over embeddings (see Tables 3 and 4), particularly in the CD-based approach,

where isolated models performed better with TF-IDF. This outcome may be attributed to the lower complexity of the classification method, a trend also observed in the fusion strategies. Specifically, Type A, which employs a simpler approach, outperformed the more complex Type B strategy. There, three out of five machine learning models achieved better performance than more structured and computationally intensive models. Regarding performance (see Figures 6 and 7), this phenomenon is evident, as the Type A fusion achieved better ROC curves compared to other approaches.

The use of NLP and LLMs in TB diagnosis has been explored in various contexts. These include cough signal analysis and embeddings for detection [49], as well as a feature fusion approach that combines cough signals with spectrogram images [50], achieving a sensitivity of 98.13% from 144 signals. Additionally, text analysis has been used to differentiate TB from COVID-19 based on reported cases [51]. Other studies have compared NLP-based image analysis with radiologist-generated natural text from chest radiographs (CXRs) [34] and developed a report generation system for lung diseases by fusing text reports with CXR images, achieving an accuracy of 94% on a dataset of 3955 images [52]. Furthermore, laboratory test results, tumor markers, and imaging analysis were used to classify 296 patients with spinal TB and spinal tumors using ChatGPT (Version 4) and ML models. The results showed that ChatGPT's performance was suboptimal, with a sensitivity of 71.67% [53]. While similar studies have explored the fusion of structured and unstructured data for TB prediction—achieving a mean AUC of 95.5% across different models [54]—the present study's approach, which focuses on basic text from a specific segment of the EMR, remains largely unexplored. Notably, the aforementioned study analyzed 692,949 patients and incorporated more advanced clinical variables (e.g., hemoptysis, cough, and erythrocyte sedimentation rate), making direct comparisons challenging. Additionally, a multimodal approach leveraging EHR datasets, various clinical notes, LLMs, and hypergraph modeling of structured EHR data was introduced by [31]. Although this method achieved AUC values of 83.54% and 73.01% on two different datasets, it was applied outside the TB domain. A pre-trained model combining text reports and CXR images achieved a validation accuracy of 94% [52], though a notable distinction was the use of approximately 3800 images from a well-established dataset. Finally, exploratory studies have leveraged NLP to extract TB-related information from EMRs [38,55], yet none have implemented classification for diagnostic support. These findings underscore the potential for future research to refine and expand this field further.

In the specific context of PTB diagnosis, clinical signs and symptoms serve as the initial indicators, identifying a patient as a potential case. According to the Colombian national protocol, at least one of the three diagnostic tests outlined in the introduction is required for confirmation. Smear microscopy, which involves analyzing sputum samples under a microscope to detect Mycobacterium, has a sensitivity of only 40–60% due to the high bacterial load needed for detection [56].

The present results, particularly the models based on embeddings and Type B data fusion, achieved sensitivity values of 73.6% and 71.3%, respectively. These models could serve as viable alternatives when only smear microscopy is available, assisting healthcare professionals in patient assessment. A more sensitive test, such as culture in solid or liquid media, requires at least two weeks and specialized infrastructure [57], which may not be feasible in low-income settings. The most advantageous option in terms of speed and accuracy—the molecular test—is also largely inaccessible in Colombia, where fewer than twenty certified facilities can perform it [58]. Furthermore, reliance on sample transportation presents logistical challenges that may compromise test integrity. Given these constraints, improving infrastructure for advanced diagnostic methods remains a

priority in developing countries like Colombia. Meanwhile, the present proposal could provide valuable clinical insights, supporting timely diagnosis and treatment initiation.

The main limitation of this study is the dataset size, which impacts the ability to train robust ML models. In AI applications, larger datasets generally lead to better performance. However, the data used in this study were obtained from a real-world setting at Hospital Santa Clara, one of the institutions treating the highest number of PTB patients. Integrating data from multiple healthcare centers and consolidating records prior to confirmed diagnoses remains a significant challenge, even within the same city. Additionally, access to data with consistent characteristics is constrained by national protocols for reporting TB cases.

Moreover, while the advancements in LLM architectures, particularly Transformers, offer promising avenues for text-based medical analysis, they were beyond the scope of this study. Future research could explore these architectures for deeper insights. However, a critical challenge with such approaches is the need for a significantly larger dataset to properly train and fine-tune the models. Expanding the dataset depends on the epidemiological characteristics of the disease and the availability of data from medical centers. Although this study's dataset is relatively small, it represents the most comprehensive PTB patient data available from Bogotá, D.C.'s leading treatment center.

4. Conclusions

The aim of this study is to explore computational intelligence tools as alternatives for providing health professionals with additional information on PTB diagnosis. The proposed approach is particularly useful in decision-making scenarios where resources are limited, and standard diagnostic tests are not readily available. This work leverages two sources of information: clinical data and physicians' reports extracted from restricted patient medical data. Various ML models were applied to each data source, along with two data fusion approaches based on text representation.

Models that relied solely on physicians' reports performed 9% worse than those using clinical data but achieved higher sensitivity (73%). While data fusion models did not surpass the performance of models based solely on clinical variables, they still demonstrated notable sensitivity. However, as discussed, clinical variables are not always accessible, and specific constraints must be considered. In such cases, NLP techniques play a crucial role by offering an alternative when patient data are incomplete or when individuals are unwilling to disclose personal information. Both fusion approaches achieved AUC values of 69.3% and 67.2%, outperforming smear microscopy, which remains the primary diagnostic option in resource-limited regions. Additionally, the embedding technique for text vectorization yielded lower performance compared to TF-IDF, which achieved an AUC of 60.8%. This suggests that TF-IDF may be a more effective method for representing textual data in the context of TB diagnosis.

The insights gained and the algorithms developed in this study contribute to future research efforts aimed at refining these approaches. Ultimately, this work lays the foundation for integrating NLP-driven tools into the workflow of healthcare professionals to enhance PTB diagnosis. As a future direction, there is an opportunity to analyze additional data sources that could enhance diagnostic accuracy. Incorporating laboratory test results, clinical examination findings, and other accessible diagnostic methods could provide valuable insights, particularly in remote regions with limited healthcare infrastructure. Additionally, when available, CXR images could serve as a complementary tool for improving TB detection and supporting clinical decision-making.

Author Contributions: Conceptualization, A.D.O.-C. and A.F.R.-G.; methodology, A.D.O.-C. and A.F.R.-G.; validation, A.D.O.-C., A.F.R.-G., A.L.J., C.E.A., M.A.P. and E.V.; formal analysis, A.D.O.-C., A.F.R.-G., A.L.J., C.E.A., M.A.P. and E.V.; investigation, A.D.O.-C., A.F.R.-G., A.L.J., C.E.A., M.A.P. and E.V.; resources, C.E.A., M.A.P. and E.V.; data curation, A.D.O.-C., A.F.R.-G. and A.L.J.; writing—original draft preparation, A.D.O.-C. and A.F.R.-G.; writing—review and editing, A.D.O.-C. and A.L.J.; visualization, A.D.O.-C. and A.F.R.-G.; supervision, A.D.O.-C., A.L.J., C.E.A., M.A.P. and E.V.; project administration, A.L.J. and A.D.O.-C.; funding acquisition, A.D.O.-C., E.V. and C.E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MinCiencias (*Ministerio de Ciencia, Tecnología e Innovación de Colombia*) grant number 123380762899 CT 819-2018, Subred Integrada de Servicios de Salud Centro-Oriente E.S.E, and Universidad Antonio Nariño, and the APC was funded by Universidad del Rosario.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Subred Integrada de Servicios de Salud Centro-Oriente E.S.E with protocol code CEI. 0106/2021, act number 316, and date of approval 27 May 2021.

Informed Consent Statement: Informed consent was not required, as all data were retrospective and fully anonymized.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The authors acknowledge the support of Universidad del Rosario, Universidad Antonio Nariño and the Subred Integrada de Servicios de Salud Centro-Oriente E.S.E in this project. Additionally, they extend their gratitude to the research seedbed team Semillero en Inteligencia Artificial en Salud (Semill-IAS) for their contributions. The authors also recognize the support of the Ministerio de Ciencia y Tecnología–Minciencias of Colombia. Finally, ChatGPT-3.5 was used as a complementary tool for reviewing language structure, grammar, and spelling.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. *Global Tuberculosis Report 2024*; World Health Organization: Geneva, Switzerland, 2024.
2. TB/COVID-19 Global Study Group. Tuberculosis and COVID-19 co-infection: Description of the global cohort. *Eur. Respir. J.* **2022**, *59*, 2102538. [CrossRef] [PubMed]
3. 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. *Eur. Rev. Med. Pharmacol. Sci.* **2022**, *26*, 2692–2701. [PubMed]
4. Shimoda, M.; Yoshiyama, T.; Okumura, M.; Tanaka, Y.; Morimoto, K.; Kokutou, H.; Osawa, T.; Furuuchi, K.; Fujiwara, K.; Ito, K.; et al. Analysis of risk factors for pulmonary tuberculosis with persistent severe inflammation: An observational study. *Medicine* **2022**, *101*, e29297. [CrossRef] [PubMed]
5. Va not 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 Syst. Rev.* **2022**, *3*, CD010890. [CrossRef]
6. 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. *J. Med. Case Rep.* **2021**, *15*, 1–7. [CrossRef]
7. Programa Nacional de Prevención y Control de la Tuberculosis, Ministerio de Salud. *Informe de Evento Tuberculosis Año 2024*; Technical Report; Programa Nacional de Prevención y Control de la Tuberculosis, Ministerio de Salud: Bogota, Colombia, 2024.
8. Ministerio de Salud y Protección Social. *Resolución Número 0000227 de 2020*; Technical Report; Ministerio de Salud y Protección Social: Bogota, Colombia, 2020.
9. 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.
10. Ministerio de Salud y Protección Social. *Plan Estratégico: Hacia el fin de la Tuberculosis, Bogota, Colombia 2016–2025*; Technical Report; Ministerio de Salud y Protección Social: Bogota, Colombia, 2016.

11. Orjuela-Cañón, A.D.; Jutinico, A.L.; Awad, C.; Vergara, E.; Palencia, A. Machine learning in the loop for tuberculosis diagnosis support. *Front. Public Health* **2022**, *10*, 876949. [CrossRef]
12. 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. *Comput. Methods Programs Biomed.* **2018**, *157*, 11–17. [CrossRef]
13. Orjuela-Cañón, A.D.; Jutinico, A.L.; González, M.E.D.; García, C.E.A.; Vergara, E.; Palencia, M.A. Time series forecasting for tuberculosis incidence employing neural network models. *Heliyon* **2022**, *8*, e09897. [CrossRef]
14. 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 Digit. Med.* **2020**, *3*, 17. [CrossRef]
15. Rajan, S.P.; Paranthaman, M. Artificial Intelligence in Healthcare: Algorithms and Decision Support Systems. In *Smart Systems for Industrial Applications*; Wiley-Scrivener: Beverly, MA, USA, 2022; pp. 173–197.
16. Shortliffe, E.H.; Sepúlveda, M.J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **2018**, *320*, 2199–2200. [CrossRef] [PubMed]
17. Jamshidi, M.B.; Daneshfar, F. A Hybrid Echo State Network for Hypercomplex Pattern Recognition, Classification, and Big Data Analysis. In Proceedings of the 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 17–18 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7–12.
18. Amisha; Malik, P.; Pathania, M.; Rathaur, V.K. Overview of artificial intelligence in medicine. *J. Fam. Med. Prim. Care* **2019**, *8*, 2328–2331.
19. Jamshidi, M.B.; Ebadpour, M.; Moghani, M.M. Cancer Digital Twins in Metaverse. In Proceedings of the 2022 20th International Conference on Mechatronics-Mechatronika (ME), Pilsen, Czech Republic, 7–9 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
20. Jamshidi, M.; Lalbakhsh, A.; Talla, J.; Peroutka, Z.; Hadjilooei, F.; Lalbakhsh, P.; Jamshidi, M.; La Spada, L.; Mirmozafari, M.; Dehghani, M.; et al. Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment. *IEEE Access* **2020**, *8*, 109581–109595. [CrossRef] [PubMed]
21. 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. In Proceedings of the 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 1–4 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 30–34.
22. Orjuela-Cañon, A.D.; Perdomo, O. Clustering proposal support for the COVID-19 making decision process in a data demanding scenario. *IEEE Lat. Am. Trans.* **2021**, *19*, 1041–1049. [CrossRef]
23. Campos, M.S.R.; Rodríguez, D.C.; Orjuela-Cañón, A.D. Tuberculosis Drug Discovery Estimation Process by Using Machine and Deep Learning Models. In *Applications of Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 43–53.
24. Dande, P.; Samant, P. Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review. *Tuberculosis* **2018**, *108*, 1–9. [CrossRef]
25. Orjuela-Cañón, A.D.; de Seixas, J. Fuzzy-ART neural networks for triage in pleural tuberculosis. In Proceedings of the 2013 Pan American Health Care Exchanges (PAHCE), Medellin, Colombia, 29 April–4 May 2013; pp. 1–4.
26. 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. *Int. J. Med. Inform.* **2018**, *112*, 149–157. [CrossRef]
27. 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. [CrossRef]
28. Wang, Y.; Luo, J.; Hao, S.; Xu, H.; Shin, A.Y.; Jin, B.; Liu, R.; Deng, X.; Wang, L.; Zheng, L.; 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. [CrossRef]
29. Cai, T.; Giannopoulos, A.A.; Yu, S.; Kelil, T.; Ripley, B.; Kumamaru, K.K.; Rybicki, F.J.; Mitsouras, D. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics* **2016**, *36*, 176–191. [CrossRef]
30. Doan, S.; Conway, M.; Phuong, T.M.; Ohno–Machado, L. Natural language processing in biomedicine: A unified system architecture overview. *Methods Mol. Biol.* **2014**, *1168*, 275–294.
31. 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* **2024**, arXiv:2403.08818.
32. Goel, A.; Gueta, A.; Gilon, O.; Liu, C.; Erell, S.; Nguyen, L.H.; Hao, X.; Jaber, B.; Reddy, S.; Kartha, R.; et al. LLMs accelerate annotation for medical information extraction. In Proceedings of the Machine Learning for Health (ML4H), New Orleans, LA, USA, 10 December 2023; PMLR: Birmingham, UK, 2023; pp. 82–100.
33. 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* **2024**, arXiv:2402.10153.
34. 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. *Clin. Imaging* **2022**, *87*, 34–37.
35. Zhao, F.; Zhang, C.; Geng, B. Deep Multimodal Data Fusion. *ACM Comput. Surv.* **2024**, *56*, 1–36. [CrossRef]
36. Pineda Rincón, E.A.; Moreno-Sandoval, L.G. Design of an architecture contributing to the protection and privacy of the data associated with the electronic health record. *Information* **2021**, *12*, 313. [CrossRef]

37. Potdar, K.; Pardawala, T.S.; Pai, C.D. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int. J. Comput. Appl.* **2017**, *175*, 7–9. [CrossRef]
38. 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. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 2468–2471.
39. Zhu, W.; Zhang, W.; Li, G.Z.; He, C.; Zhang, L. A study of damp-heat syndrome classification using Word2vec and TF-IDF. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1415–1420.
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Egger, R. Text Representations and Word Embeddings. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*; Springer International Publishing: Cham, Switzerland, 2022; pp. 335–361.
42. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
43. 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. *Sci. Rep.* **2022**, *12*, 6256. [CrossRef]
44. 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. *Semin. Diagn. Pathol.* **2023**, *40*, 71–87.
45. 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 J. Integr. Biol.* **2020**, *24*, 241–246. [CrossRef]
46. 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: Birmingham, UK, 2015.
47. Keras. 2015. Available online: https://github.com/fchollet/keras (accessed on 2 May 2022).
48. 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 Fail.* **2019**, *6*, 428–435. [CrossRef] [PubMed]
49. Pahar, M.; Theron, G.; Niesler, T. Automatic Tuberculosis detection in cough patterns using NLP-style cough embeddings. In Proceedings of the 2022 International Conference on Engineering and Emerging Technologies (ICEET), Virtual, 27–28 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
50. Xu, W.; Bao, X.; Lou, X.; Liu, X.; Chen, Y.; Zhao, X.; Zhang, C.; Pan, C.; Liu, W.; Liu, F. Feature fusion method for pulmonary tuberculosis patient detection based on cough sound. *PLoS ONE* **2024**, *19*, e0302651. [CrossRef] [PubMed]
51. Pholo, M.D.; Hamam, Y.; Khalaf, A.B.; Du, C. Differentiating Between COVID-19 and Tuberculosis Using Machine Learning and Natural Language Processing. *Rev. d'Intell. Artif.* **2022**, *36*, 313–318. [CrossRef]
52. Naz, I.; Iftikhar, S.; Zahra, A.; Zainab, S. Report Generation of Lungs Diseases from Chest X-Ray Using NLP. *Int. J. Innov. Sci. Technol.* **2022**, *3*, 223–233.
53. Hu, X.; Xu, D.; Zhang, H.; Tang, M.; Gao, Q. Comparative diagnostic accuracy of ChatGPT-4 and machine learning in differentiating spinal tuberculosis and spinal tumors. *Spine J.* 2025, *in press*. [CrossRef]
54. 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 Min.* **2023**, *16*, 11. [CrossRef]
55. Landsman, D.; Abdelbasit, A.; Wang, C.; Guerzhoy, M.; Joshi, U.; Mathew, S.; Pou-Prom, C.; Dai, D.; Pequegnat, V.; Murray, J.; et al. Cohort profile: St. Michael's Hospital Tuberculosis Database (SMH-TB), a retrospective cohort of electronic health record data and variables extracted using natural language processing. *PLoS ONE* **2021**, *16*, e0247872. [CrossRef]
56. 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.; et al. 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. *Clin. Infect. Dis.* **2017**, *64*, e1–e33. [CrossRef]
57. Ghazvini, K.; Yousefi, M.; Firoozeh, F.; Mansouri, S. Predictors of tuberculosis: Application of a logistic regression model. *Gene Rep.* **2019**, *17*, 100527. [CrossRef]
58. 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. [CrossRef]