Comparison of Tree-Based Machine Learning Models for Classification of Tuberculosis Outcomes in Brazil

Heloísa de Almeida Pereira, Marcos Roberto Ribeiro, and Ciniro Aparecido Leite Nametala

***Abstract*—Tuberculosis remains a significant public health concern, recognized as a reemerging disease strongly associated with socioeconomic conditions. According to the World Health Organization, tuberculosis continues to be the leading cause of death from a single infectious agent worldwide in 2025. This study evaluates Random Forest, XGBoost, CatBoost, and LightGBM for classifying four treatment outcomes (cure, abandonment, death from tuberculosis, and death from other causes) using 53,656 epidemiological records from Minas Gerais, Brazil (2010–2024). A preprocessing pipeline was designed to handle heterogeneous and high-cardinality variables. Models were assessed with and without SMOTE balancing under 5-fold stratified cross-validation, using accuracy, weighted F1-score, and per-class recall as evaluation metrics. Without balancing, all models achieved approximately 0.75 accuracy but failed to detect minority outcomes. Experiments across multiple random seeds, feature subset analysis, and temporal validation were conducted to assess model robustness. With SMOTE, CatBoost achieved the best overall performance, with the highest cross-validation F1-score (0.7071 ± 0.0034) and improved recall for abandonment and death outcomes. Results indicate that the combination of clinical and socioeconomic features is essential for predictive performance, and that data quality and class imbalance are the main obstacles to reliable minority-class detection in tuberculosis surveillance.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10507

***Index Terms*—Tuberculosis, machine learning, clinical outcomes.**

I. Introduction

TUBERCULOSIS remains a significant concern in public health, including in developed countries, which underscores the challenges related to its control [1]. According to the World Health Organization (WHO), tuberculosis continues to be the leading cause of death from a single infectious agent worldwide in 2025, a position temporarily occupied by COVID-19 [2]. The disease's persistence reflects the intersection of biological, social, and structural factors that challenge public health systems. Studies indicate that countries with lower per capita investment in healthcare and basic infrastructure present higher incidence rates [3]–[5].

The associate editor coordinating the review of this manuscript and approving it for publication was Anabel Martin (*Corresponding author: Heloísa de Almeida Pereira*).

Heloísa de Almeida Pereira, M. R. Ribeiro, and C. A. L. Nametala are with the Federal Institute of Minas Gerais, Brazil (e-mails: 0065523@academico.ifmg.edu.br, marcos.ribeiro@ifmg.edu.br, and ciniro.nametala@ifmg.edu.br).

In Brazil, the persistence of tuberculosis is related to the level of socioeconomic development and to failures in the organization and management of health systems [3], [6]. Minas Gerais, in particular, presents a widespread prevalence of cases, with emphasis on the central macro-region, which concentrated the highest occurrence in the state between 2015 and 2020 [6]. The occurrence of tuberculosis is driven by social determinants, affecting vulnerable populations. Among the main factors associated with late diagnosis and high disease prevalence, the absence or insufficiency of basic health services stands out [3], [6].

Recent advances in Artificial Intelligence (AI) have enabled machines to perform cognitive tasks with levels of performance comparable to, and sometimes exceeding, human capabilities [7]–[11]. As a field of computer science dedicated to creating intelligent systems, AI has expanded its role across multiple sectors, particularly in healthcare. Its applications range from resource management and clinical research support to diagnostic prediction and drug development [11], [12]. This expansion has been driven by the growing availability of computational infrastructure and the reduction of technology costs, fostering the digital transformation of healthcare systems [13]. Machine learning (ML) and deep learning (DL) techniques have been used to identify patterns in large-scale biomedical and social datasets, supporting tasks such as disease prediction, patient stratification, and optimization of treatment strategies [14]–[16].

In public health, AI plays a role in identifying patterns of social vulnerability, enabling data-driven and preventive interventions focused on reducing inequalities in access to healthcare [17]. Despite the growing body of research on AI applied to tuberculosis, existing studies present technical gaps that limit their applicability to outcome prediction. Most studies focus on tuberculosis detection using radiographic data and deep learning architectures [18], leaving treatment outcome classification underexplored, particularly in real-world epidemiological data. Existing approaches also prioritize disease detection over outcome classification and risk factor identification. Studies that address treatment outcomes rely on unidimensional socioeconomic indicators, which fail to capture the interaction between social vulnerability and clinical variables [19].

Motivated by these challenges, this study evaluates four tree-based ensemble methods in classifying tuberculosis treatment outcomes in Brazil: Random Forest, XGBoost, CatBoost, and LightGBM. The investigation addresses class im-

balance through the Synthetic Minority Oversampling Technique (SMOTE), assessing the capacity of these methods to improve the identification of minority outcomes and contribute to epidemiological surveillance. Tree-based ensemble methods suit this context because they handle heterogeneous and imbalanced tabular data, require minimal preprocessing, and provide feature importance metrics that support risk factor identification.

The main contributions of this work are:

- A comparative evaluation of Random Forest, XGBoost, CatBoost, and LightGBM for multiclass classification of tuberculosis treatment outcomes using real-world epidemiological data.
- The design of a preprocessing pipeline integrating frequency encoding, one-hot encoding, and feature standardization to handle heterogeneous and high-cardinality variables.
- An analysis of the impact of class imbalance using SMOTE, highlighting its role in improving minority class detection.
- An experimental framework based on stratified cross-validation, multiple random seeds, and ablation studies to evaluate model stability and robustness.
- An investigation of the contribution of clinical and socioeconomic features, demonstrating that their combination improves predictive performance.

II. Related Works

Tuberculosis remains a public health issue in Brazil, especially in contexts of social inequality. Freitas et al. [6] analyzed records of adult pulmonary tuberculosis in Belo Horizonte, using data from the *Sistema de Informação de Agravos de Notificação* (Notifiable Diseases Information System) [20], [21], covering the period from 2001 to 2020. The study found that mandatory variables such as age, sex, and diagnostic tests showed high completeness. However, variables for public policy planning, such as education level, race, and comorbidities, had fair or insufficient completion. The limited quality of these data weakens disease monitoring, hinders case control, and restricts the identification of priority populations. The authors highlight the need for strategies to improve data recording and strengthen epidemiological surveillance systems.

From a technological standpoint, Hansun et al. [18] conducted a systematic review to evaluate the performance of AI-based methods for tuberculosis detection. The research examined ML and DL algorithms applied to different data modalities, including radiographic, molecular, and physiological data.

A total of 152 studies published up to July 2023 were analyzed, selected from Scopus [22], PubMed [23], and ACM Digital Library [24] databases. Most works used radiographic data (84.9%) and DL techniques (80.3%), with an emphasis on convolutional neural networks such as VGG-16, ResNet-50, and DenseNet-121. The models were evaluated using standard metrics: accuracy, sensitivity, specificity, and area under the ROC curve (AUC). Reported results showed an average model accuracy of 91.9%, sensitivity 92.8%, specificity 92.4%, and mean AUC 93.5%. Models employing transfer learning showed more consistent performance and better results compared to models trained from scratch. However, only one study assessed the model robustness regarding the domain-shift problem, limiting the generalization of results to real-world data [18].

Studies combining ML and DL achieved better results in some metrics, although still underexplored. The authors noted that the predominance of radiographic data and the lack of molecular or physiological data limit multimodal analysis. Nevertheless, when available, multimodal approaches tend to generalize better under standardized evaluation. The review also highlights the importance of using explainable methods (Explainable Artificial Intelligence – XAI), such as Grad-CAM and LIME, to interpret model outputs and facilitate their adoption in clinical settings. The conclusion reinforces that AI-based methods show promising performance for tuberculosis detection, but future studies should include evaluation on real-world data, concrete clinical applications, and more diverse data, particularly from neglected populations.

Strika et al. conducted a systematic review on the application of AI in underserved regions [13], identifying three main areas of impact: assisted screening and diagnosis, epidemiological prediction and surveillance, and resource management support. Kitsios et al. [11] highlight that technologies such as deep learning algorithms, big data, and automated clinical screening systems expand the reach of healthcare services, particularly in vulnerable regions, optimizing the allocation of professionals and infrastructure.

Finally, Citron et al. [19] argue that poverty, in its multiple dimensions, influences tuberculosis incidence and outcomes. The critical analysis of socioeconomic assessment methods showed that most tuberculosis studies still use unidimensional indicators, such as income, failing to capture the complexity of individual deprivation. The authors advocate for the adoption of multidimensional poverty indices adapted to local contexts and for the standardization of data collection in epidemiological research. The use of more comprehensive indicators may help identify vulnerable groups, guide social policies, and improve interventions aimed at combating tuberculosis.

III. Methodology

The methodological design of this study was structured in six stages, as illustrated in Fig. 1. This framework was designed to evaluate how tree-based ML algorithms can enhance the classification of tuberculosis treatment outcomes and contribute to strengthening epidemiological surveillance in Minas Gerais.

This approach integrates epidemiological analysis with computational modeling, aiming to identify vulnerability patterns and assess the algorithms in detecting cases at higher risk of adverse outcomes.

The first stage consisted of a literature review, which examined healthcare inequality and access, particularly in Brazil, as well as research addressing tuberculosis and the application of AI models in public health. The objective was to identify current challenges, assess the scientific relevance of the topic, and support the theoretical foundation of the present study.

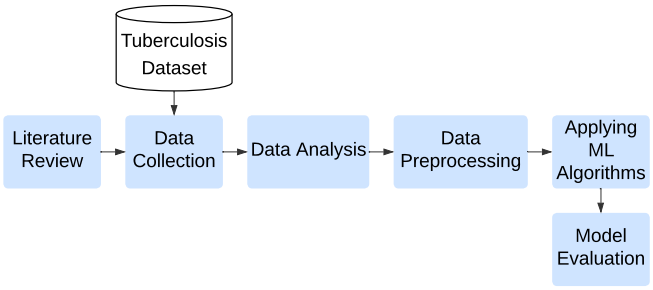

Fig. 1. Proposed methodology workflow for the study: literature review, data collection, data analysis, preprocessing, ML modeling, and evaluation.

In the second stage, data were collected from the Tuberculosis dataset, which compiles epidemiological and clinical information from confirmed cases. The third stage focused on data analysis. At this point, the completeness and consistency of the variables were evaluated, aiming to detect potential gaps or inconsistencies that might affect the quality of the results. This stage also served to validate the dataset and ensure its suitability for modeling purposes.

The fourth stage involved data preprocessing, including cleaning missing or inconsistent variables, normalizing variable formats, and converting categorical data when necessary.

Finally, the fifth and sixth stages comprised the development and evaluation of predictive models. Different ML algorithms were trained and tested using a hold-out strategy (train-test split), combined with stratified cross-validation applied to the training set, to assess their performance in classifying tuberculosis treatment outcomes. The evaluation included an ablation study to quantify the contribution of feature engineering and SMOTE to model performance, a feature subset analysis comparing clinical, socioeconomic, and complete variable sets, and a temporal analysis to assess model stability across different notification periods. A comparative analysis of results was then carried out, highlighting the models with the most reliable and consistent performance metrics.

Random Forest [27], XGBoost [28], CatBoost [29], and LightGBM [30] were selected for their robustness with heterogeneous datasets and nonlinear relationships, both common in epidemiological data. These methods also provide feature importance metrics that support the identification of clinical and sociodemographic factors associated with treatment outcomes.

SMOTE was incorporated to address the class imbalance between the majority outcome (Cure) and minority outcomes (Abandonment and Death), enabling a more balanced evaluation of model performance on less frequent but clinically critical cases.

A. Dataset

The dataset used to evaluate the occurrence of tuberculosis in Minas Gerais refers to the registry of reported cases in the state, covering the period from January 1, 2010, to April 4, 2024. This dataset was extracted from the *Portal de Dados Abertos do Estado de Minas Gerais* (Open Data Portal of the State of Minas Gerais), published on April 22, 2024 [25]. All data were obtained from official governmental sources and contain anonymized records, ensuring compliance with ethical and data protection standards.

The first part of the data analysis consisted of validating the dataset and verifying the completeness and consistency of the variables. Variables with a high proportion of missing or inconsistent values were excluded, as well as those without direct analytical relevance to the study objectives. Further details regarding the dataset structure, data dictionary, and variables prior to the selection process are available in the project repository [26]. The selected variables are shown in Table I.

TABLE I
VARIABLES IN THE TUBERCULOSIS DATASET

Variable	Description	Usage
ID_MUNICIP	Municipality of notification.	Freq. encoded
ID_REGIONA	Regional health district.	Categorical
ID_UNIDADE	Reporting health unit.	Freq. encoded
DT_DIAG	Diagnosis date.	Derived
NU_IDADE_N	Patient's age at notification.	Numerical
CS_SEXO	Biological sex.	Categorical
CS_RACA	Race or skin color.	Categorical
CS_ESCOL_N	Education level.	Categorical
ID_MN_RESI	Municipality of residence.	Freq. encoded
TRATAMENTO	Type of treatment entry.	Categorical
CULTURA_ES	Result of culture test.	Categorical
HISTOPATOL	Result of histopathology.	Categorical
DT_INIC_TR	Start date of treatment.	Derived
SITUA_ENCE	Case outcome or closure.	Target
FORMA	Clinical form (pulmonary, extrapulmonary, both).	Categorical
AGRAVAIDS	AIDS comorbidity.	Categorical
AGRAVALCOO	Alcoholism comorbidity.	Categorical
AGRAVDIABE	Diabetes comorbidity.	Categorical
AGRAVDOENC	Mental disorder comorbidity.	Categorical
AGRAVOUTRA	Other comorbidities.	Categorical
HIV	HIV test result.	Categorical

Categorical: one-hot encoded.
Freq. encoded: converted to relative frequency.
Derived: used to compute `tempo_inicio_trat`.
Numerical: used as-is after standardization.
Target: outcome variable.

The final dataset, after validation and exclusion of inconsistent records, comprises 53,656 samples and 19 variables, of which 5 are numerical and 14 are categorical. A total of 1,595 missing values were identified across the dataset, concentrated primarily in sociodemographic fields such as education level, race, and comorbidities, which are non-mandatory fields in the notification system, a pattern consistent with findings reported by Freitas et al. [6].

To further explore the dataset, a series of descriptive and exploratory analyses was performed to identify epidemiological patterns, frequency distributions, and possible associations among clinical, sociodemographic, and geographic factors. These analyses were guided by the following research questions:

1) Distribution of cases with and without comorbidities: The relative frequency of tuberculosis cases presenting comorbidities such as HIV, AIDS, alcoholism, diabetes, mental disorders. This step aimed to quantify the comorbidity burden among individuals diagnosed with tuberculosis and to explore potential relationships between the

disease and other health conditions that may influence its progression.

2) Cure rate among patients with and without comorbidities: The proportion of cases with the outcome "cure" was calculated within each comorbidity group and compared to patients without associated conditions. This comparison sought to determine whether the presence of comorbidities adversely affects treatment success rates.
3) Occurrence of deaths according to profile: The absolute and relative frequencies of deaths attributed to tuberculosis were calculated across groups with and without comorbidities. This comparison aimed to assess the effect of pre-existing health conditions on tuberculosis mortality.
4) Annual evolution of notifications between 2010 and 2023: Records from 2024 were excluded due to incomplete data availability (records up to April only). The time series enabled the visualization of temporal trends in the number of tuberculosis notifications over the 14-year period.
5) Annual percentage variation in reported cases: The year-to-year relative variation in case numbers was calculated and expressed in both absolute and percentage terms. This measure provided insight into the periods of increase and decline in tuberculosis notifications over time.
6) Distribution by educational level: The educational levels were organized into incomplete/complete primary, secondary, and higher education. The relative proportions of each category were calculated to describe the educational profile of the reported population.
7) Distribution by clinical form of tuberculosis: Classify cases as pulmonary, extrapulmonary, or mixed. This allowed evaluation of the prevalence of different clinical manifestations across the dataset.
8) Distribution by biological sex: Aiming to identify sex-related differences in tuberculosis occurrence.
9) Distribution by age group: The relative frequency of each age group was computed to identify which demographic segments exhibited the highest number of reported cases.

All these descriptive analyses contributed to a clearer and more comprehensive understanding of the population affected by tuberculosis. They enabled the identification of key epidemiological patterns and highlight existing sociodemographic disparities. The insights gained from this stage served as a foundation for the following phases of the study, which focused on applying ML models.

B. Proposed Architecture

Fig. 2 shows the structure adopted for tuberculosis outcome classification. The architecture defines the data flow from preprocessing to model evaluation.

The process starts with the input dataset and applies text normalization, missing data imputation, feature transformation, categorical encoding, and numerical standardization.

A stratified hold-out split separates training and test sets. The training set supports model development under a Stratified K-Fold cross-validation scheme, with SMOTE applied only to the training partitions.

The test set remains unchanged and supports final evaluation. All models follow the same processing structure.

C. Implementation Platform

All experiments were conducted in Python [31] within the Jupyter Notebook environment [32], on a local workstation equipped with an AMD Ryzen 7 4800H processor (2.90 GHz, 8 cores) and 16 GB of RAM. Data manipulation and preprocessing were performed using `pandas` [33], `NumPy` [34], `unidecode` [35], and `openpyxl` [36]. Pipeline construction, cross-validation, and performance metrics were handled by `scikit-learn` [37], with additional statistical support from `scipy` [38]. Class imbalance was addressed using the `imbalanced-learn` library [39]. Exploratory data analysis and visualization were performed using `matplotlib` [40], `seaborn` [41], and `folium` [42] for geospatial mapping.

The evaluated models were implemented using their respective libraries: `scikit-learn` for Random Forest, `xgboost` [28], `catboost` [29], and `lightgbm` [30]. The complete list of libraries and their respective versions is available in the project repository [26].

D. Preprocessing

Data preprocessing was performed in multiple steps to ensure that the tuberculosis dataset was consistent, standardized, and ready for modeling. First, the text in categorical variables was normalized by removing accents, trimming unnecessary spaces, and replacing blank spaces with underscores. Subsequently, only cases belonging to the relevant treatment outcomes were selected: Cure, Death from tuberculosis, Death from other causes, and Abandonment. Records outside this range were excluded.

Next, the date variables (`DT_DIAG` and `DT_INIC_TR`) were converted into a standard datetime format. From them, a new variable was derived: the time to treatment initiation, calculated as the number of days between diagnosis and the start of treatment. Missing values in this variable were imputed using the median to mitigate the impact of outliers, and duplicate entries were removed to avoid redundancy.

For categorical variables with a large number of unique values, such as municipality (`ID_MUNICIP`), health unit (`ID_UNIDADE`), and place of residence (`ID_MN_RESI`), frequency encoding was applied. This approach converted each category into its relative frequency within the dataset, helping models handle these high-cardinality variables efficiently. Variables with fewer unique categories, including sex, race, education level, clinical form, and presence of comorbidities, were preserved for later encoding.

Subsequently, a hold-out strategy was adopted, and the dataset was divided into training and testing subsets using an 80/20 stratified split, ensuring that the class distribution was preserved across both sets. This resulted in 42,924 samples for training and 10,732 samples for testing, with consistent proportions in each subset.

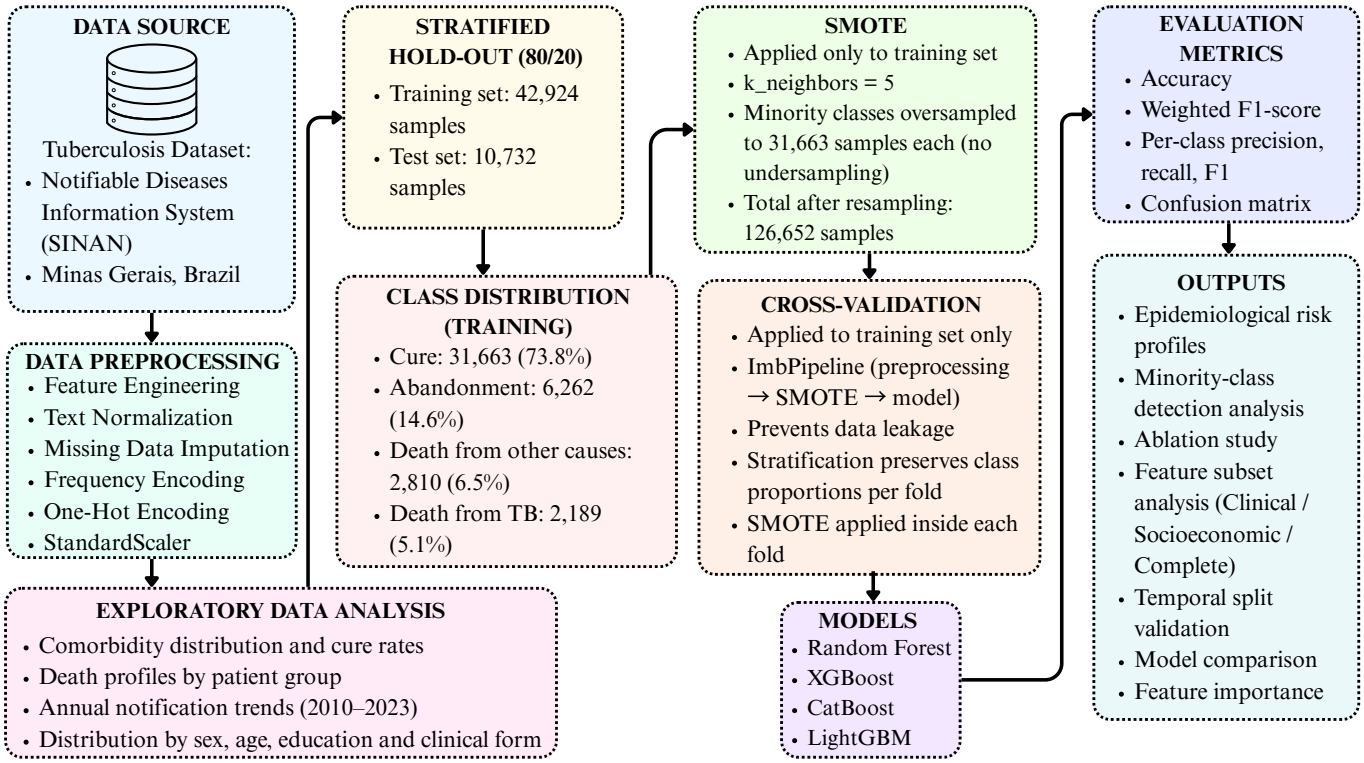

Fig. 2. Proposed architecture for the study: from data source and preprocessing through exploratory analysis, stratified hold-out split, SMOTE-based cross-validation, model training, and evaluation.

In the training set, Cure accounts for 31,663 cases (73.77%), followed by Abandonment with 6,262 cases (14.59%), Death from other causes with 2,810 cases (6.55%), and Death from tuberculosis with 2,189 cases (5.10%). A similar distribution is observed in the test set, with 7,917 cases (73.77%) for Cure, 1,566 (14.59%) for Abandonment, 702 (6.54%) for Death from other causes, and 547 (5.10%) for Death from tuberculosis, as shown in Fig. 3.

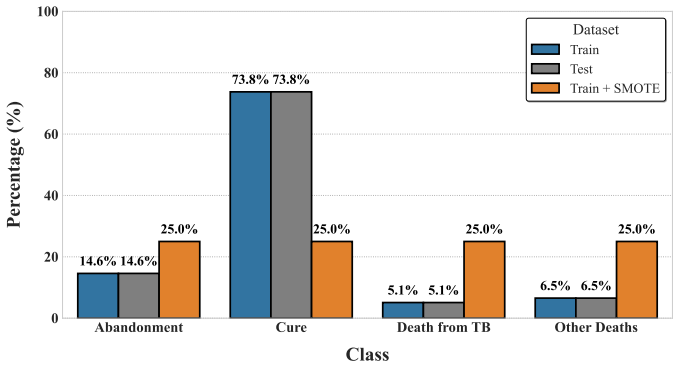

Fig. 3. Percentage distribution of tuberculosis outcome classes across the original dataset, training and test partitions, and the training set after SMOTE-based balancing.

Distribution analyses of all variables are available in the project repository [26]. Fig. 4 presents the probability density functions of patient age and treatment initiation delay. The age distribution approximates a normal curve, with mean 43.8 years and median 43.0 years. The treatment initiation delay distribution is right-skewed, with median 0.0 days and mean 2.0 days, indicating that most patients began treatment without delay.

To address the predominance of the Cure class, SMOTE was applied exclusively to the training data, using `k_neighbors=5`. After resampling, all classes became equally represented with 31,663 samples each (25.00%), totalling 126,652 training samples. No undersampling of the majority class was performed.

For feature transformation, a ColumnTransformer was defined to process categorical and numerical variables in parallel. Categorical variables with low cardinality were encoded using One-Hot Encoding, converting each category into binary features. Numerical variables, including age, time to treatment initiation, and frequency-encoded attributes, were standardized using StandardScaler, ensuring zero mean and unit variance. These transformation steps were incorporated into a Pipeline structure together with the classifier, ensuring that all transformations are learned exclusively from the training data and applied consistently during evaluation. For experiments involving class balancing, an ImbPipeline was used to include SMOTE between preprocessing and model training, guaranteeing that oversampling is performed only on transformed training data and preventing data leakage.

A Stratified K-Fold cross-validation strategy ($k = 5$) was applied to the training set. In each fold, the complete pipeline: preprocessing, optional SMOTE, and model training, was fitted on the training partition and evaluated on the validation partition. Stratification ensures that class proportions are preserved

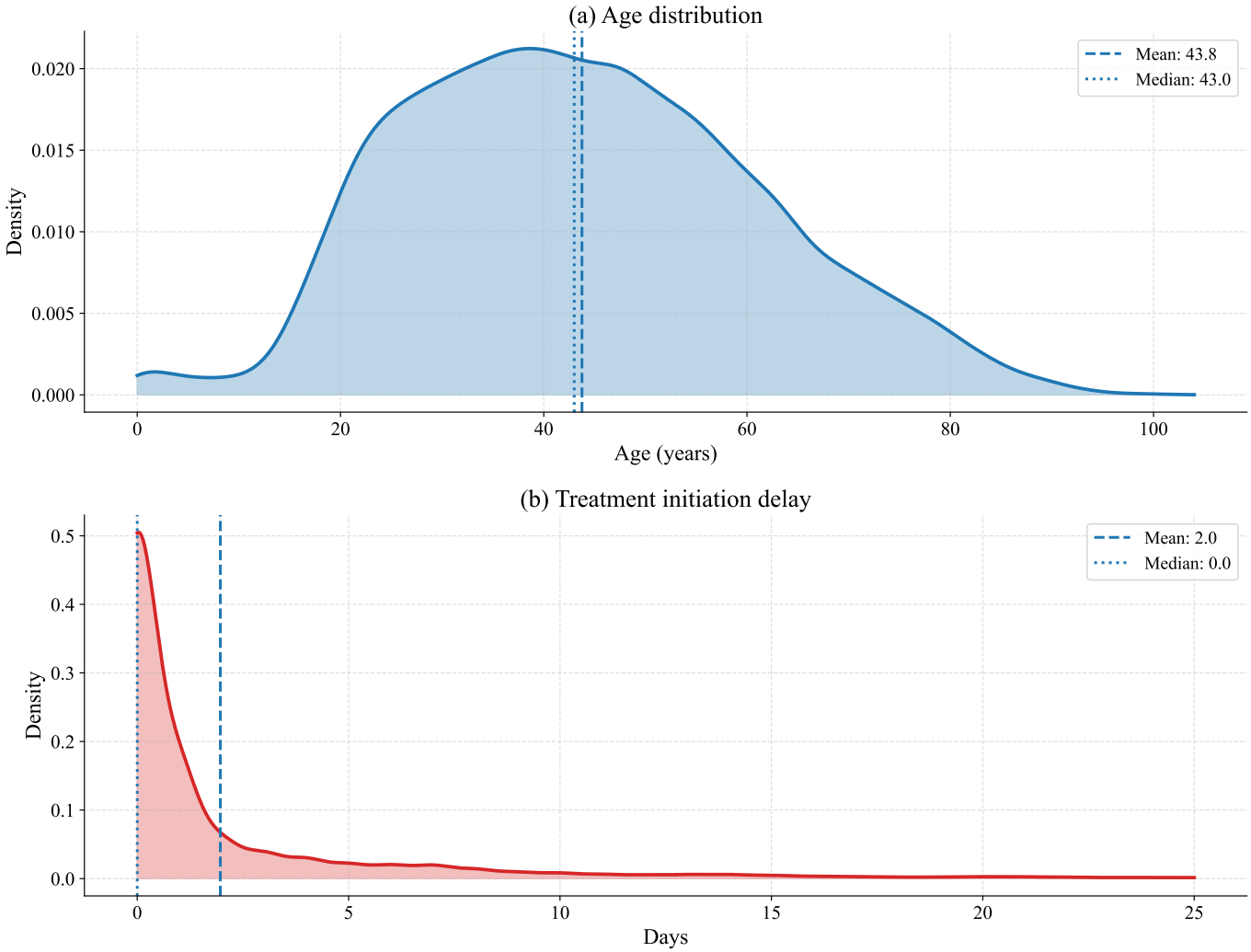

Fig. 4. Probability density functions of the continuous variables analyzed in this study: (a) age distribution, with mean and median values of 43.8 and 43.0 years, respectively; and (b) treatment initiation delay (days), showing a right-skewed distribution with mean and median values of 2.0 and 0.0 days. Dashed and dotted vertical lines indicate the mean and median, respectively.

across all folds, providing a more stable estimate of model performance.

E. Model Evaluation

In the modeling stage, Random Forest, XGBoost, CatBoost, and LightGBM were used to classify treatment outcomes for tuberculosis patients. Each model was tested under two distinct conditions: (1) using the original imbalanced data distribution and (2) applying SMOTE to mitigate class imbalance by generating synthetic samples for minority classes.

The outcome variable was encoded using LabelEncoder to ensure numerical compatibility across all models, without altering class semantics.

The hyperparameters used for each model are summarized in Table II. Random Forest was configured with 200 decision trees and balanced class weights. XGBoost was set with 300 estimators, a learning rate of 0.1, and a maximum depth of 6, using mlogloss as the evaluation metric. CatBoost was defined with 300 iterations, a learning rate of 0.1, and a depth of 6, adopting MultiClass as the loss function. Finally, LightGBM was configured with 300 estimators, a learning rate of 0.1, and no explicit depth limit (max_depth = -1), allowing trees to grow according to data-driven constraints. A multiclass objective with balanced class weights was adopted to handle class imbalance.

TABLE II
HYPERPARAMETERS USED IN THE EVALUATED MODELS

Model	Parameter	Value
Random Forest	n_estimators	200
	class_weight	balanced
XGBoost	n_estimators	300
	learning_rate	0.1
	max_depth	6
CatBoost	iterations	300
	learning_rate	0.1
	depth	6
LightGBM	n_estimators	300
	learning_rate	0.1
	max_depth	-1
	class_weight	balanced

Random Forest builds an ensemble of decision trees, each trained on a bootstrap sample of the dataset. At each node split, the algorithm considers only a random subset of features, reducing correlation among trees and controlling variance. This study uses 200 estimators and sets `class_weight=balanced`, which adjusts sample weights inversely proportional to class frequencies, compensating for the predominance of the Cure class without resampling.

XGBoost builds trees sequentially, each fitted to the negative gradient of the loss function from the previous iteration. For multiclass problems, the loss function corresponds to softmax cross-entropy (mlogloss). The learning rate of 0.1 controls the contribution of each tree, while the maximum depth of 6 constrains tree complexity to reduce overfitting. In addition, XGBoost incorporates regularization and subsampling mechanisms that improve generalization and robustness when handling structured tabular data.

CatBoost processes categorical features through ordered target statistics computed on permuted subsets of the training data, so the encoding of each sample depends only on preceding observations. This ordered boosting strategy reduces the risk of target leakage and prediction shift. The configuration uses 300 iterations, learning rate 0.1, depth 6, and the MultiClass loss function with softmax output, directly optimizing the model for the four-class task.

LightGBM grows trees leaf-wise, expanding the leaf with the highest loss reduction at each step. Two mechanisms distinguish it from other boosting implementations: Gradient-based One-Side Sampling (GOSS), which retains instances with large gradients and subsamples those with small gradients, and Exclusive Feature Bundling (EFB), which reduces dimensionality by merging mutually exclusive sparse features. Setting `max_depth=-1` allows the algorithm to determine tree depth from data, which may contribute to its sensitivity to class distribution and the large performance gap observed between the unbalanced and SMOTE conditions.

The four algorithms represent distinct approaches to ensemble learning: Random Forest reduces variance through bagging and averaging; XGBoost and CatBoost reduce bias through sequential boosting with regularization; LightGBM prioritizes computational efficiency through histogram approximations and selective sampling. All four produce feature importance scores and handle nonlinear interactions in tabular data, which meets the interpretability and robustness requirements of epidemiological classification tasks.

The hyperparameters used in this study were defined through a manual tuning process. Different configurations were empirically tested for each model, and their performance was evaluated based on cross-validation results and classification metrics. These values follow common choices reported in tabular data literature. XGBoost, CatBoost, and LightGBM used a learning rate of 0.1, a default value that controls the step size at each iteration, shrinking the contribution of each new tree while providing stable convergence. XGBoost and CatBoost used a maximum depth of 6, a common default that limits tree complexity while capturing interactions among predictor variables. LightGBM kept max_depth = -1, letting the leaf-wise growth strategy determine tree structure directly from the data.

The number of estimators balanced predictive performance and computational cost. Random Forest used 200 trees; the boosting algorithms used 300. These configurations produced stable results, with standard deviations below 0.005 across folds and random seeds. The use of class_weight=balanced in Random Forest and LightGBM served as a baseline strategy for handling class imbalance, allowing the contribution of SMOTE to be isolated from algorithm-level weighting. This study did not perform systematic hyperparameter optimization (grid search or Bayesian optimization with nested cross-validation); results compare the algorithms under a common set of empirically chosen configurations rather than individually optimized settings.

The models were evaluated using accuracy, precision, recall, and weighted F1-score, with confusion matrices generated to provide a detailed view of class-specific performance and reveal misclassification patterns. In addition to the main evaluation, a cross-validation analysis was conducted on the training set to assess model stability and generalization. An ablation study was performed to quantify the contribution of individual preprocessing steps, and a feature subset analysis compared three configurations: clinical variables only, socioeconomic variables only, and the complete feature set. A temporal split validation was also carried out to assess model robustness under a chronological data partition, simulating a more realistic deployment scenario. Beyond the computational evaluation, the results were examined for their practical implications in public health, given that the early identification of abandonment or death risk is essential to guide targeted interventions and optimize healthcare resource allocation.

IV. Results and Discussion

This section presents and discusses the performance results obtained from the ML models applied to tuberculosis treatment outcomes. The classification task considered four possible outcomes: Abandonment (0), Cure (1), Death from tuberculosis (2), and Death from other causes (3).

A. Data Analysis

The graphs and detailed analyses supporting the findings described in this section are available in the project repository [26].

The data analysis identified clear social and clinical disparities. Abandonment and death were concentrated in specific patient profiles. The highest proportions occurred among men aged 25–44 years, representing the economically active group. Individuals with low education levels, especially those with only primary education or no formal schooling, exhibited higher abandonment rates. These results indicate that socioeconomic conditions directly influence treatment continuity, consistent with the findings reported by Queiroz et al. and Biggs et al. [3], [5].

Alcohol use disorder was one of the strongest factors associated with abandonment. Patients with a history of alcoholism were more likely to discontinue therapy. Additionally, mental health disorders were frequently observed, often co-occurring

with alcohol use, which further compromised treatment adherence.

Deaths were mainly associated with HIV co-infection and delayed treatment initiation. Patients who started therapy more than 60 days after diagnosis showed higher mortality. Among HIV-positive individuals, deaths accounted for a larger proportion of total outcomes compared with other groups.

Both abandonment and death were more frequent among patients with lower education levels. These same groups also had higher rates of missing information, especially in education and occupation variables, which limits the analysis and the model's ability to identify risk patterns.

Overall, the results indicate that social vulnerability, comorbidities, and delayed treatment are the key factors associated with abandonment and death in tuberculosis cases.

Informed by these findings, models were trained to classify tuberculosis treatment outcomes. However, because abandonment and death are infrequent and often linked to incomplete socioeconomic data, the dataset remained inherently unbalanced. This imbalance impacted model performance, causing algorithms to prioritize the majority class (Cure) and reducing sensitivity for minority outcomes. The following subsections detail the performance of each model, emphasizing their capacity to detect abandonment and death.

B. Random Forest

Random Forest achieved an accuracy of 0.7461, with high recall for Cure (0.9704) and low recall for the minority classes: Abandonment (0.1596), Death from tuberculosis (0.0402), and Death from other causes (0.0741), as shown in Table III. The confusion matrix shows that most abandonment and death cases were misclassified as cure, reflecting the influence of class imbalance (Fig. 5).

TABLE III
RANDOM FOREST - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.5230	0.1596	0.2446
1	0.7621	0.9704	0.8537
2	0.5116	0.0402	0.0746
3	0.4031	0.0741	0.1252
Accuracy			0.7461

To ensure a consistent interpretation in the multiclass setting, the elements of the confusion matrix were computed using a one-vs-rest approach. For each outcome, the class of interest was considered the positive class, while the remaining outcomes were grouped as negative. Thus, for a given outcome (e.g., abandonment), correctly predicted cases of that outcome were counted as true positives, cases incorrectly assigned to that outcome were considered false positives, cases of that outcome predicted as other categories were treated as false negatives, and all remaining correctly classified instances were considered true negatives. This procedure was consistently applied to all outcomes, namely cure, abandonment, death from tuberculosis, and death from other causes.

After applying SMOTE, as detailed in Table IV, accuracy decreased to 0.7229, but recall improved for the minority classes (abandonment: 0.2433, death from tuberculosis: 0.1261, death from other causes: 0.1994).

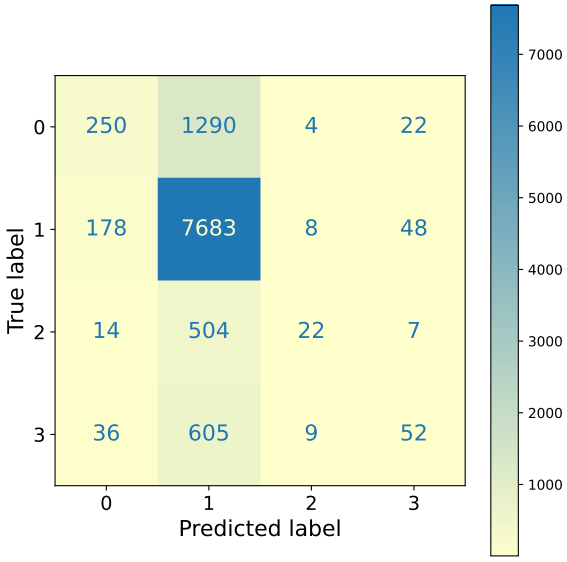

Fig. 5. Confusion matrix of the Random Forest model for the four tuberculosis outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

TABLE IV
RANDOM FOREST WITH SMOTE - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.3965	0.2433	0.3015
1	0.7863	0.9054	0.8417
2	0.3350	0.1261	0.1833
3	0.3118	0.1994	0.2433
Accuracy			0.7229

The confusion matrix shows more true positives for critical outcomes (Fig. 6). This indicates that balancing partially corrected the bias toward the majority class, at the cost of higher misclassification among cured cases.

In summary, Random Forest captured the general structure of the dataset but remained limited by the imbalance and missing socioeconomic data, particularly affecting predictions of abandonment and death.

C. XGBoost

XGBoost achieved 0.7486 accuracy and high recall for Cure (0.9596). Recall for Abandonment (0.1839) and the two death outcomes (0.1225 and 0.1168) remained low (Table V). The confusion matrix shows the same concentration of predictions in the majority class (Fig. 7).

With SMOTE, accuracy slightly decreased to 0.7395, but recall improved for abandonment (0.2369), death from tuberculosis (0.1335), and death from other causes (0.2051) (Table VI). The model preserved precision while improving minority-class detection, showing good adaptation to oversampled data (Fig. 8).

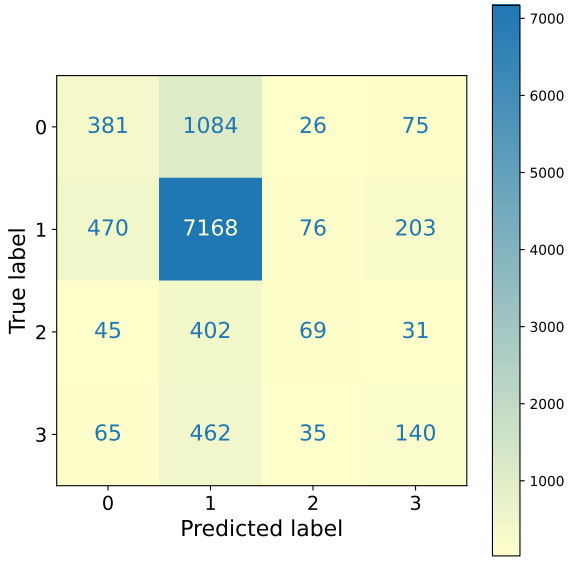

Fig. 6. Confusion matrix of the Random Forest model trained on the SMOTE-balanced dataset for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

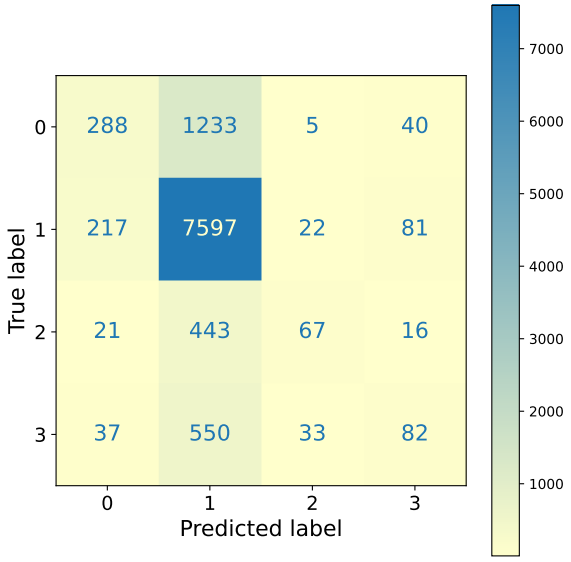

Fig. 7. Confusion matrix of the XGBoost model for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

TABLE V
XGBOOST - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.5115	0.1839	0.2705
1	0.7734	0.9596	0.8565
2	0.5276	0.1225	0.1988
3	0.3744	0.1168	0.1781
Accuracy			0.7486

D. CatBoost

CatBoost without SMOTE reached an accuracy of 0.7522, with recall of 0.9697 for Cure and 0.1775, 0.0878, and 0.0997 for the three minority classes (Table VII). As in the previous models, most abandonment and death cases were predicted as Cure. (Fig. 9).

With SMOTE, accuracy decreased to 0.7301, while recall for the minority classes increased (abandonment: 0.2656, death from tuberculosis: 0.1463, death from other causes: 0.2308) (Table VIII). The confusion matrix (Fig. 10) confirms this redistribution. CatBoost also showed stable weighted averages, indicating consistency across different resampling strategies.

Overall, CatBoost achieved a better compromise between classes, showing balanced recall values while maintaining general accuracy, making it one of the most reliable models in this analysis.

TABLE VI
XGBOOST WITH SMOTE - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.4673	0.2369	0.3144
1	0.7851	0.9281	0.8507
2	0.3946	0.1335	0.1995
3	0.3655	0.2051	0.2628
Accuracy			0.7395

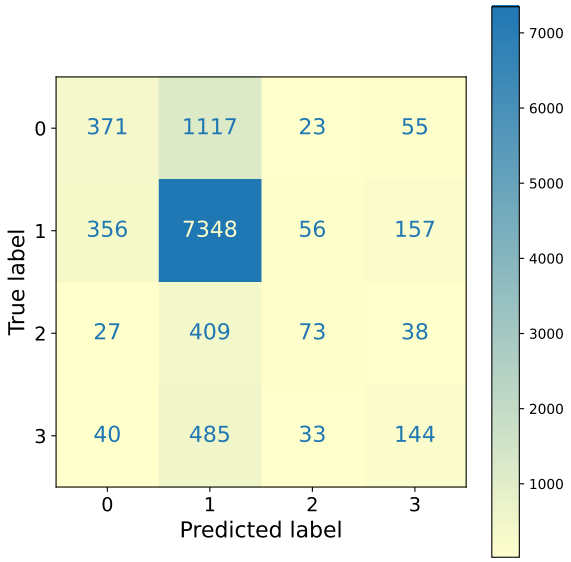

Fig. 8. Confusion matrix of the XGBoost model trained on the SMOTE-balanced dataset for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

E. LightGBM

LightGBM exhibited performance distinct from the other models. Without SMOTE, accuracy was 0.5905, with low recall for Cure (0.6355) and higher recall for the death classes (0.4406, 0.4330) (Table IX).

The confusion matrix (Fig. 11) reveals significant misclassification across classes, indicating that the model struggled to separate boundaries effectively under strong imbalance, likely due to aggressive class weighting.

TABLE VII
CATBOOST - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.5505	0.1775	0.2685
1	0.7698	0.9697	0.8582
2	0.5854	0.0878	0.1526
3	0.4070	0.0997	0.1602
Accuracy			0.7522

TABLE VIII
CATBOOST WITH SMOTE - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.4411	0.2656	0.3316
1	0.7929	0.9065	0.8459
2	0.3306	0.1463	0.2028
3	0.3273	0.2308	0.2707
Accuracy			0.7301

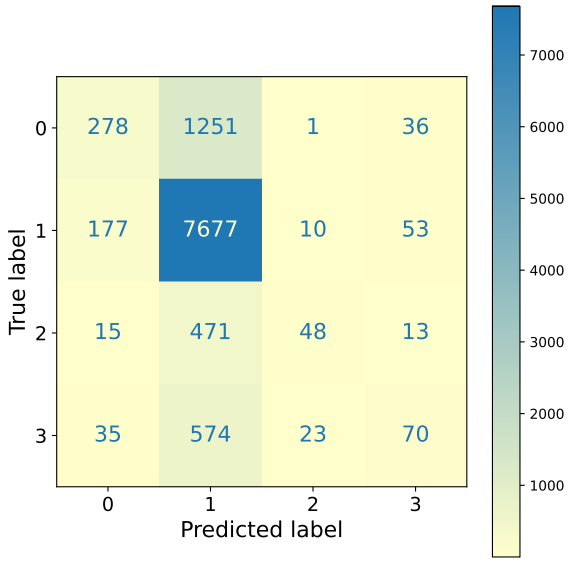

Fig. 9. Confusion matrix of the CatBoost model for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

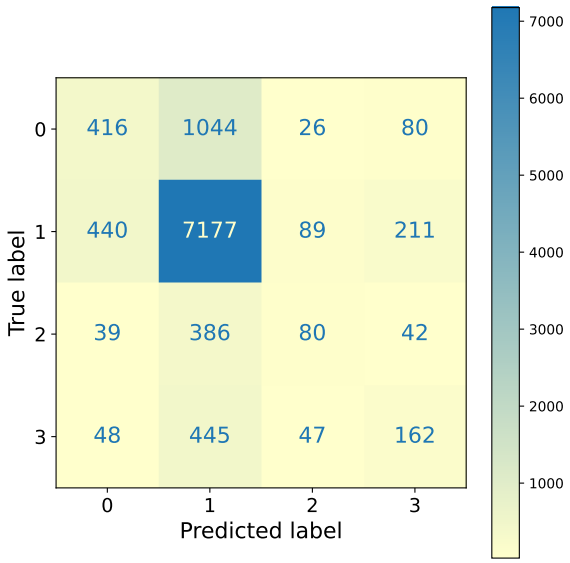

Fig. 10. Confusion matrix of the CatBoost model trained on the SMOTE-balanced dataset for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

After applying SMOTE, accuracy increased to 0.7432, and recall improved for Cure (0.9406), while also improving overall reliability (Table X). This suggests that LightGBM is more sensitive to resampling methods, as shown in Fig. 11, and benefits substantially from balanced training data.

This behavior indicates that LightGBM depends strongly on data balancing to achieve reliable performance across all outcomes.

F. Ablation Study

An ablation study was conducted using a Random Forest classifier as a fixed baseline model to evaluate the contribution of feature engineering and data balancing with SMOTE. By keeping the model unchanged, the analysis isolates the impact of data representation and class imbalance handling on predictive performance. Experiments were repeated with multiple random seeds to evaluate performance variability.

Table XI presents the results in terms of cross-validation performance and test performance averaged over five different seeds.

Feature engineering produced a small increase in performance compared to the baseline. In contrast, the addition of SMOTE resulted in a larger improvement, indicating the impact of class imbalance on the task.

Additionally, the low standard deviation across cross-validation and test runs indicates stable performance across different random initializations.

G. Feature Subsets

Using the same Random Forest configuration, experiments were conducted with different subsets of input variables: (i) clinical features, (ii) socioeconomic features, and (iii) the complete feature set, as shown in Table XII.

The results indicate that both clinical and socioeconomic features contribute to predictive performance, but neither is sufficient alone. Models trained with only clinical variables showed the lowest performance, suggesting that clinical information alone does not fully capture the factors associated with treatment outcomes.

Socioeconomic features yielded better results, reinforcing the influence of social vulnerability identified in the data analysis. However, the best performance was achieved using the complete feature set, indicating that combining both sources of information is essential. These results reinforce findings from the theoretical background: social vulnerability and clinical variables interact in ways that unidimensional indicators fail to capture [18], [19].

TABLE IX
LIGHTGBM - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.3240	0.4860	0.3888
1	0.8685	0.6355	0.7339
2	0.1966	0.4406	0.2719
3	0.2229	0.4330	0.2943
Accuracy			0.5905

TABLE X
LIGHTGBM WITH SMOTE - CLASSIFICATION REPORT

Class	Precision	Recall	F1-score
0	0.4769	0.2171	0.2984
1	0.7814	0.9406	0.8537
2	0.4400	0.1207	0.1894
3	0.3628	0.1752	0.2363
Accuracy			0.7432

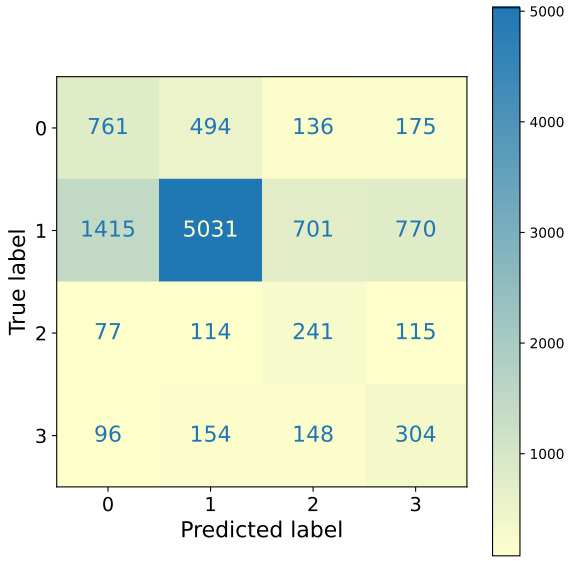

Fig. 11. Confusion matrix of the LightGBM model for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

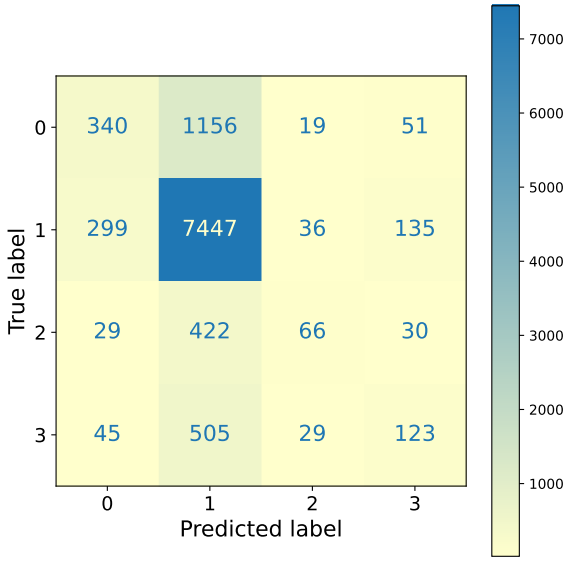

Fig. 12. Confusion matrix of the LightGBM model trained on the SMOTE-balanced dataset for the four tuberculosis treatment outcomes: 0 - Abandonment, 1 - Cure, 2 - Death from TB, and 3 - Death from other causes.

H. Temporal Analysis

To evaluate model generalization under realistic temporal conditions, a walk-forward validation strategy was applied. At each iteration, the model was trained on all records prior to year t and tested on records from year t, with t ranging over the available years in the dataset. Years with fewer than 30 test records were excluded. SMOTE was applied exclusively to the training set within each fold to prevent data leakage.

Fig. 13 presents the weighted F1-score and per-class recall across test years.

Performance varied across years, reflecting distributional shifts in the data over time. The weighted F1-score remained generally below the values obtained in the random split setting, consistent with the increased difficulty of temporal generalization, with a more pronounced decline observed in the most recent years (2022–2024). Recall for Cure remained the highest across all years. Recall for Abandonment showed moderate but unstable values, while recall for Death from tuberculosis and Death from other causes remained consistently low throughout the evaluation period, indicating limited model sensitivity to these rare outcomes under temporal shift. These results confirm that class imbalance and data heterogeneity across years constitute the main obstacles to reliable minority-class detection in a prospective deployment scenario.

This decline in the most recent years reflects three factors. First, the COVID-19 pandemic disrupted tuberculosis surveillance in Brazil during 2020–2022: case notifications dropped 13% in 2020 and 9% in 2021, and studies estimate around 20,000 TB cases went undiagnosed nationally between April 2020 and December 2021 [43], [44]. This disruption created a temporal distribution shift between the training data from earlier years and the 2022–2024 test sets. Second, the walk-forward scheme uses a cumulative training window, so the model progressively incorporates data from different reporting regimes, which reduces its sensitivity to the most recent patterns.

Third, the low prevalence of Death from tuberculosis and Death from other causes makes yearly recall estimates highly sensitive to small variations in case counts under a temporal split. This explains the noisy and consistently low recall curves for these classes. Prospective deployment requires periodic retraining and targeted monitoring of minority-class recall, since aggregate metrics such as accuracy or weighted F1-score can mask clinically important degradation in rare-outcome detection.

I. Model Comparison

Tables XIII and XIV summarize the comparative performance of all evaluated models. The metrics include weighted precision, weighted recall, weighted F1-score, and overall accuracy, which represent the average performance across all

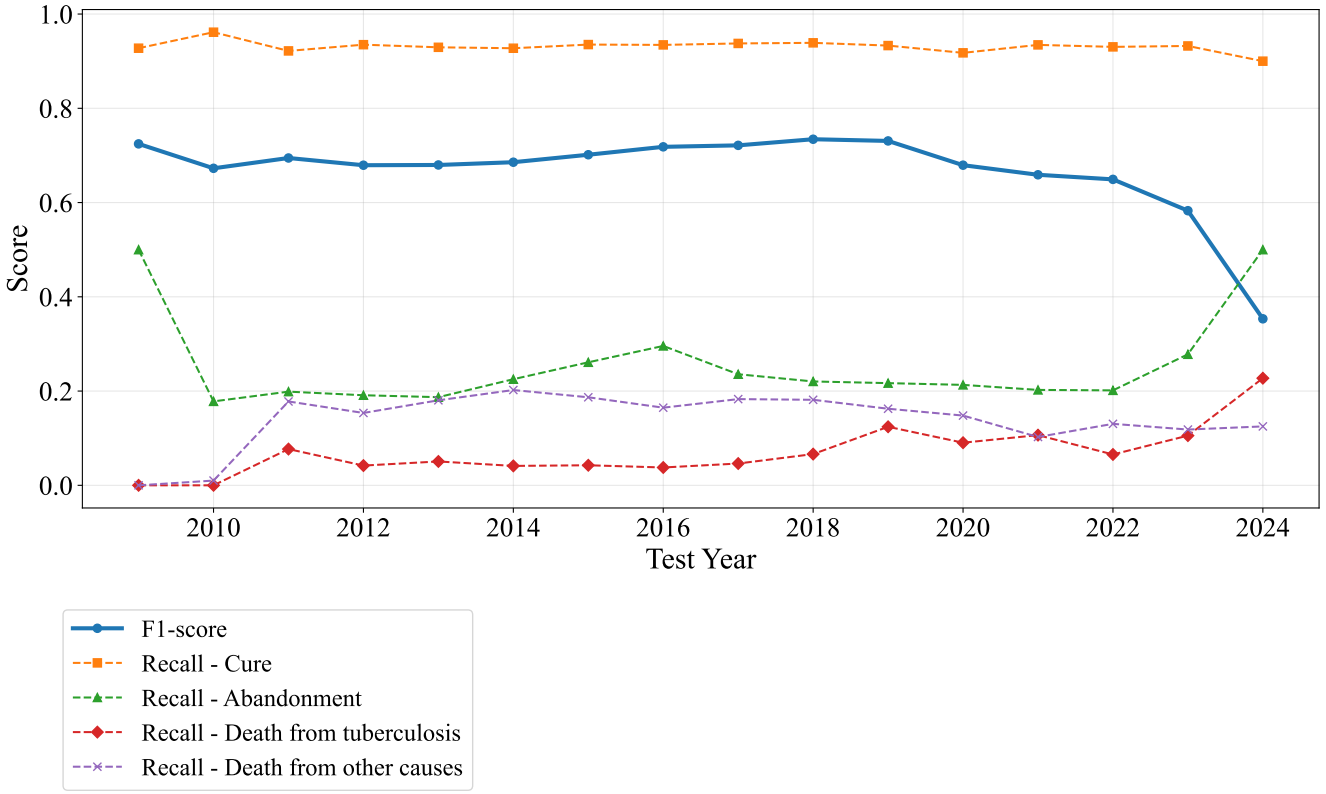

Fig. 13. Walk-forward temporal validation: weighted F1-score and per-class recall across test years.

TABLE XI
ABLATION STUDY: IMPACT OF FEATURE ENGINEERING AND SMOTE ON WEIGHTED F1-SCORE (RANDOM FOREST, FIVE RANDOM SEEDS)

Config.	CV F1	Std	95% CI	Test F1	Std
Baseline	0.6745	0.0026	(0.671, 0.678)	0.6707	0.0016
+ FE	0.6767	0.0025	(0.674, 0.680)	0.6774	0.0028
+ FE + SMOTE	**0.6940**	0.0036	(0.690, 0.699)	**0.6912**	0.0027

TABLE XII
WEIGHTED F1-SCORE BY FEATURE SUBSET (CLINICAL, SOCIOECONOMIC, COMPLETE) USING RANDOM FOREST WITH SMOTE

Subset	CV F1	Std	95% CI	Test F1	Std
Clinical	0.6254	0.0030	(0.622, 0.629)	0.6208	0.0042
Socioeconomic	0.6431	0.0046	(0.637, 0.649)	0.6433	0.0023
Complete	**0.6933**	0.0024	(0.690, 0.696)	**0.6923**	0.0027

TABLE XIII
WEIGHTED PRECISION, WEIGHTED RECALL, F1-SCORE, AND ACCURACY OF ALL MODELS WITHOUT SMOTE

Model	W. Prec.	W. Rec.	W. F1	Acc.
Random Forest	0.6977	0.7475	0.6791	0.7475
XGBoost	0.6965	0.7495	**0.6944**	0.7495
CatBoost	0.7074	**0.7530**	0.6925	**0.7530**
LightGBM	**0.7077**	0.5852	0.6260	0.5852

TABLE XIV
WEIGHTED PRECISION, WEIGHTED RECALL, F1-SCORE, AND ACCURACY OF ALL MODELS WITH SMOTE

Model	W. Prec.	W. Rec.	W. F1	Acc.
Random Forest	0.6771	0.7238	0.6914	0.7238
XGBoost	**0.6924**	0.7401	**0.7013**	0.7401
CatBoost	0.6892	0.7308	**0.7013**	0.7308
LightGBM	0.6910	**0.7423**	0.6989	**0.7423**

outcome classes while accounting for their unequal frequencies in the dataset.

To further assess model robustness, a 5-fold cross-validation procedure was conducted. Table XV presents the mean weighted F1-score, standard deviation, and 95% confidence intervals for each model.

Cross-validation results indicate low variability across folds, with standard deviations below 0.005 for all models, confirming stable and consistent performance. Confidence intervals were narrow, reinforcing the reliability of the estimates.

Across both test and cross-validation results, XGBoost and CatBoost achieved the best balance between accuracy and recall. CatBoost with SMOTE obtained the highest mean F1-score, followed by XGBoost and LightGBM with SMOTE. Random Forest showed consistent but lower sensitivity to minority outcomes, while LightGBM demonstrated the largest

TABLE XV
5-FOLD CROSS-VALIDATION PERFORMANCE

Model	Mean F1	Std	95% CI
Random Forest	0.6780	0.0028	(0.6741, 0.6819)
Random Forest + SMOTE	0.6930	0.0039	(0.6876, 0.6984)
XGBoost	0.6976	0.0027	(0.6939, 0.7014)
XGBoost + SMOTE	0.7048	0.0029	(0.7007, 0.7088)
CatBoost	0.6933	0.0042	(0.6874, 0.6992)
CatBoost + SMOTE	**0.7071**	0.0031	(0.7028, 0.7113)
LightGBM	0.6353	0.0016	(0.6331, 0.6376)
LightGBM + SMOTE	0.7025	0.0012	(0.7008, 0.7042)

improvement after applying SMOTE, indicating strong dependence on data balancing.

Overall, the results demonstrate that model performance is influenced more by data characteristics than by the choice of algorithm. Class imbalance, missing information, and feature representation play a central role in predictive performance.

The extended evaluation framework, including multi-seed experiments, feature subset analysis, and temporal validation, reveals that combining clinical and socioeconomic data is essential for improving generalization. Furthermore, while SMOTE consistently improves the detection of minority outcomes, predicting rare events such as abandonment and death remains a challenging task, particularly under real-world temporal shifts.

CONCLUSIONS

This study evaluated the potential of tree-based ML models to classify tuberculosis treatment outcomes in Minas Gerais and examined how data imbalance and socioeconomic disparities shape predictive performance. The exploratory analysis showed that abandonment and death are concentrated in specific groups, especially individuals with low education, alcohol use disorder, HIV co-infection, and delayed treatment initiation. These patterns highlight structural vulnerabilities that directly influence outcome distribution within the dataset.

Across all models, performance was affected by the predominance of the Cure class and by missing information in key socioeconomic variables. Without balancing, the algorithms achieved high recall for the majority class but low sensitivity for abandonment and death. After applying SMOTE, recall for minority outcomes improved, albeit at the expense of reduced overall accuracy. CatBoost and XGBoost exhibited the most stable results, while LightGBM showed the largest improvement after oversampling.

From our perspective, cross-validation results provide the most reliable basis for comparing models, as they reflect performance across multiple data partitions rather than a single split. CatBoost with SMOTE achieved the highest mean weighted F1-score of 0.7071 (95% CI: 0.7028–0.7113), followed by XGBoost with SMOTE at 0.7048 (95% CI: 0.7007–0.7088) and LightGBM with SMOTE at 0.7025 (95% CI: 0.7008–0.7042). Standard deviations remained below 0.005 across all configurations, confirming stable generalization across folds. Without SMOTE, LightGBM yielded a mean F1-score of only 0.6353, the lowest among all configurations, which indicates that its leaf-wise growth strategy requires balanced training data to function reliably. CatBoost maintained consistent performance under both conditions, reflecting the advantage of its ordered boosting strategy in handling heterogeneous and imbalanced tabular data. The ablation study reinforces this interpretation: SMOTE produced a larger F1-score gain than feature engineering alone, raising the Random Forest baseline from 0.6767 to 0.6940, which shows that correcting class imbalance contributes more to predictive performance than encoding refinements when the outcome distribution is heavily concentrated in one class. The feature subset analysis confirms that clinical and socioeconomic variables each contribute independently: the complete feature set reached a cross-validation F1-score of 0.6933, against 0.6254 for clinical-only and 0.6431 for socioeconomic-only configurations.

The findings suggest that algorithmic performance relies not only on model choice but also on the quality and completeness of epidemiological data. The models were able to capture general patterns but remained limited in identifying high-risk cases, reflecting gaps in the underlying dataset. Improving data completeness, especially for education, occupation, and behavioral variables, may enhance the detection of vulnerable groups and strengthen predictive modeling.

These findings should be interpreted in light of limitations related to data quality and study design. Most studies reported in the literature are based on radiographic data and deep learning approaches, which rely on more standardized inputs and more complete datasets. In contrast, this study uses real-world epidemiological surveillance data, which are characterized by heterogeneity, class imbalance, and incomplete recording of key variables. Variables relevant to public health planning, including education level, race, and comorbidities, presented fair or insufficient completion, which weakens disease monitoring and restricts the identification of priority populations. Consequently, although the evaluated models achieved reasonable performance in distinguishing cure outcomes, their sensitivity to treatment abandonment and death remained constrained by the underlying data structure rather than by the modeling approach.

In summary, ML can support tuberculosis surveillance by indicating which factors contribute to unfavorable outcomes and by assisting in early risk identification. However, the current results underscore the critical need for improved data quality and targeted data collection strategies to fully realize this potential. Future work should explore alternative sampling methods, cost-sensitive learning, and the integration of spatial or temporal features to improve model sensitivity to abandonment and death.

REFERENCES

[1] C. Busatto, D. V. Bierhals, J. S. Vianna, P. E. A. Silva, L. G. Possuelo, and I. B. Ramis, "Epidemiology and control strategies for tuberculosis in countries with the largest prison populations," *Rev. Soc. Bras. Med. Trop.*, vol. 55, e0060-2022, 2022, doi: https://doi.org/10.1590/0037-8682-0060-2022

[2] World Health Organization, *Global Tuberculosis Report 2025*. Geneva, Switzerland: World Health Organization, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240116924

[3] J. R. Queiroz *et al.*, "Tendência da mortalidade por tuberculose e relação com o índice sociodemográfico no Brasil entre 2005–2019," *Cien. Saude Colet.*, vol. 29, e00532023, 2024, doi: https://doi.org/10.1590/1413-81232024295.00532023

[4] J. Chakaya, M. Khan, F. Ntoumi, E. Aklillu, R. Fatima, P. Mwaba, N. Kapata, S. Mfinanga, S. E. Hasnain, P. D. M. C. Katoto, A. N. H. Bulabula, N. A. Sam-Agudu, J. B. Nachega, S. Tiberi, T. D. McHugh, I. Abubakar, and A. Zumla, "Global Tuberculosis Report 2020—Reflections on the global TB burden, treatment and prevention efforts," *Int. J. Infect. Dis.*, vol. 113, suppl. 1, pp. S7–S12, 2021, doi: https://doi.org/10.1016/j.ijid.2021.02.107

[5] B. Biggs, L. King, S. Basu, and D. Stuckler, "Is wealthier always healthier? The impact of national income level, inequality, and poverty on public health in Latin America," *Soc. Sci. Med.*, vol. 71, no. 2, pp. 266–273, 2010, doi: https://doi.org/10.1016/j.socscimed.2010.04.002

[6] G. Freitas *et al.*, "Completude dos dados de notificação e de acompanhamento da tuberculose pulmonar em Belo Horizonte, Minas Gerais, de 2001 a 2020: estudo transversal," *REME Rev. Min. Enferm.*, vol. 28, p. 1563, 2025, doi: 10.35699/2316-9389.2025.49258

[7] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion models, and transformers," *IEEE Access*, vol. 12, pp. 69812–69837, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3397775

[8] L. Schut, N. Tomašev, T. McGrath, D. Hassabis, U. Paquet, and B. Kim, "Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 122, no. 13, e2406675122, 2025, doi: https://doi.org/10.1073/pnas.2406675122

[9] M. Shin, J. Kim, B. van Opheusden, and T. L. Griffiths, "Superhuman artificial intelligence can improve human decision-making by increasing novelty," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 12, e2214840120, 2023, doi: https://doi.org/10.1073/pnas.2214840120

[10] E. Sanches, F. Augusto, P. Silveira, B. Feijó Junqueira, D. A. M. Lemes, J. Picolo, G. Ribeiro Sales, V. Corso, and C. S. Bezerra, "Artificial intelligence for gas leak detection with thermal cameras and metal oxide semiconductor sensors," *Rev. Inform. Teor. Apl.*, vol. 32, no. 1, pp. 91–98, 2025, doi: https://doi.org/10.22456/2175-2745.143526

[11] F. Kitsios, M. Kamariotou, A. I. Syngelakis, and M. A. Talias, "Recent advances of artificial intelligence in healthcare: A systematic literature review," *Appl. Sci.*, vol. 13, no. 13, 7479, 2023, doi: https://doi.org/10.3390/app13137479

[12] A. Ahmadi and N. Rabie Nezhad Ganji, "AI-driven medical innovations: Transforming healthcare through data intelligence," *Int. J. BioLife Sci.*, vol. 2, no. 2, pp. 132–142, 2023, doi: https://doi.org/10.22034/ijbls.2023.185475

[13] Z. Strika, K. Petkovic, R. Likic, and R. Batenburg, "Bridging healthcare gaps: A scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts," *Postgrad. Med. J.*, vol. 101, no. 1191, pp. 4–16, 2024, doi: https://doi.org/10.1093/postmj/qgae122

[14] L. Marques, B. Costa, M. Pereira, A. Silva, J. Santos, L. Saldanha, I. Silva, P. Magalhães, S. Schmidt, and N. Vale, "Advancing precision medicine: A review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare," *Pharmaceutics*, vol. 16, no. 3, 332, 2024, doi: https://doi.org/10.3390/pharmaceutics16030332

[15] H. Taherdoost and A. Ghofrani, "AI's role in revolutionizing personalized medicine by reshaping pharmacogenomics and drug therapy," *Intell. Pharmacy*, vol. 2, no. 5, pp. 643–650, 2024, doi: https://doi.org/10.1016/j.ipha.2024.08.005

[16] T. A. Vicentini and L. C. Bertucci Barbosa, "Machine learning applied to biological sequence comparison: An alignment-free approach," *Rev. Inform. Teor. Apl.*, vol. 32, no. 2, pp. 22–35, 2025, doi: https://doi.org/10.22456/2175-2745.141585

[17] M. Bonal, C. Padilla, G. Chevillard, and V. Lucas-Gabrielli, "A French classification to describe medical deserts: A multi-professional approach based on the first contact with the healthcare system," *Int. J. Health Geogr.*, vol. 23, no. 1, 5, 2024, doi: https://doi.org/10.1186/s12942-024-00366-7

[18] S. Hansun, A. Argha, I. Bakhshayeshi, A. Wicaksana, H. Alinejad-Rokny, G. J. Fox, S.-T. Liaw, B. G. Celler, and G. B. Marks, "Diagnostic performance of artificial intelligence-based methods for tuberculosis detection: Systematic review," *J. Med. Internet Res.*, vol. 27, e69068, 2025, doi: https://doi.org/10.2196/69068

[19] C. Cintron, M. Dauphinais, X. Du, A. Tabackman, A. Lenart, A. Laliberte, J. Dirksen, and P. Sinha, "Enriching tuberculosis research by measuring poverty better: A perspective," *BMC Glob. Public Health*, vol. 3, p. 17, 2025, doi: https://doi.org/10.1186/s44263-025-00127-z

[20] Ministério da Saúde do Brasil, *Sistema de Informação de Agravos de Notificação (SINAN)*. Brasília, Brazil: Ministério da Saúde, 2024. [Online]. Available: https://portalsinan.saude.gov.br/ [Accessed: Jan. 08, 2026].

[21] Brasil, Ministério da Saúde, *DATASUS: Departamento de Informática do Sistema Único de Saúde*. Brasília, Brazil: Ministério da Saúde, 2025. [Online]. Available: https://datasus.saude.gov.br/ [Accessed: Jan. 08, 2026].

[22] Elsevier, *Scopus: Abstract and Citation Database*. Amsterdam, Netherlands: Elsevier, 2024. [Online]. Available: https://www.scopus.com [Accessed: Jan. 08, 2026].

[23] U.S. National Library of Medicine, *Montgomery County X-ray Set*. Bethesda, MD, USA: National Library of Medicine, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/ [Accessed: Jan. 08, 2026].

[24] Association for Computing Machinery, *ACM Digital Library*. New York, NY, USA: ACM, 2024. [Online]. Available: https://dl.acm.org [Accessed: Jan. 08, 2026].

[25] Minas Gerais, Governo do Estado; Fundação de Amparo à Pesquisa do Estado de Minas Gerais; and Universidade Federal de Minas Gerais, Centro de Desenvolvimento e Planejamento Regional, *DataViva: Plataforma de visualização de dados socioeconômicos*. Belo Horizonte, Brazil, 2025. [Online]. Available: https://dataviva.info/ [Accessed: Jan. 08, 2026].

[26] H. de Almeida Pereira, M. Roberto Ribeiro, and C. Aparecido Leite Nametala, "Comparison of Tree-Based Machine Learning Models for Classification of Tuberculosis Outcomes in Brazil," Zenodo, v1.0.0, 2025, doi: https://doi.org/10.5281/zenodo.20012729

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794, doi: https://doi.org/10.1145/2939672.2939785

[29] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," arXiv:1810.11363, 2018. doi: https://doi.org/10.48550/arXiv.1810.11363

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3149–3157.

[31] G. Van Rossum and F. L. Drake Jr., *Python Reference Manual*, Python Software Foundation, 2009. Available: https://www.python.org.

[32] T. Kluyver et al., "Jupyter Notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, IOS Press, 2016, pp. 87–90. doi: https://doi.org/10.3233/978-1-61499-649-1-87

[33] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proc. 9th Python in Science Conf. (SciPy)*, 2010, pp. 56–61. doi: https://doi.org/10.25080/Majora-92bf1922-00a

[34] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020. doi: https://doi.org/10.1038/s41586-020-2649-2

[35] T. Solc, *Unidecode: ASCII Transliterations of Unicode Text*, 2024. Available: https://github.com/avian2/unidecode

[36] E. Gazoni and C. Clark, *openpyxl: A Python Library to Read/Write Excel 2010 xlsx/xlsm Files*, 2024. Available: https://openpyxl.readthedocs.io

[37] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[38] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020, doi: https://doi.org/10.1038/s41592-019-0686-2

[39] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curated List of Selection Mechanisms for Data Sampling," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017. Available: http://jmlr.org/papers/v18/16-365.html

[40] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007. doi: https://doi.org/10.1109/MCSE.2007.55

[41] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021, doi: https://doi.org/10.21105/joss.03021

[42] R. Story et al., *folium: Python Data, Leaflet.js Maps*, 2013. Available: https://github.com/python-visualization/folium

[43] T. A. A. Pontes, F. Fernandez-Llimos, and A. Wiens, "Impact of COVID-19 on tuberculosis notifications," *Rev. Inst. Med. Trop. Sao Paulo*, vol. 66, e37, 2024, doi: https://doi.org/10.1590/S1678-9946202466037

[44] M. H. Chitwood, N. A. Menzies, P. Bartholomay, D. M. Pelissari, J. N. B. Silva Junior, L. O. Harada, F. D. C. Johansen, E. L. N. Maciel, M. C. Castro, M. Sanchez, J. L. Warren, and T. Cohen, "Quantifying disruptions to tuberculosis case detection in Brazilian states during the COVID-19 pandemic," *Int. J. Epidemiol.*, vol. 54, no. 5, dyaf146, 2025, doi: https://doi.org/10.1093/ije/dyaf146

Heloísa de Almeida Pereira is a final-year undergraduate student in Computer Engineering at the Federal Institute of Minas Gerais (IFMG), Brazil, graduating in 2026. Her academic interests focus on data analysis, machine learning, and the application of computational techniques to health-related datasets, with emphasis on public health data. Her work is oriented toward the use of artificial intelligence to support data-driven analysis and research in this domain.

Marcos Roberto Ribeiro holds a Ph.D. degree and an M.Sc. degree in Computer Science from the Federal University of Uberlândia (UFU), Brazil, as well as a B.Sc. degree in Computer Science from Centro Universitário de Formiga (UNIFOR-MG), Brazil. He is a full-time Professor at the Federal Institute of Minas Gerais (IFMG) – Campus Bambuí, Brazil. His current research interests are focused on artificial intelligence, database systems, and data analysis.

Ciniro Aparecido Leite Nametala holds a Ph.D. in Electrical Engineering from the University of São Paulo (USP), Brazil, an M.Sc. in Electrical Engineering from the Federal University of Minas Gerais (UFMG), Brazil, a specialization in Software Engineering from the Federal University of Lavras (UFLA), Brazil, and a degree in Analysis and Systems Development from the Federal Institute of Minas Gerais (IFMG), Brazil. He is a full-time professor with exclusive dedication at IFMG – Campus Bambuí and a member of the Research Group on Computational Systems. His current research interests are focused on data analysis and machine learning applied to public health datasets.