ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 1, pp. 133-153
DOI:10.33168/JSMS.2023.0108

Tuberculosis Treatment Failure Classification based on Electronic Medical Records Using PCA-ANN

Sfenrianto [1], Gilang Al Qarana [2], Leonov Rianto [2]

[1] Information System Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta 11480, Indonesia

[2] IKIFA College of Health Science, East Jakarta, Jakarta, Indonesia, E-mail: gilang.alqarana@ikifa.ac.id (Corresponding Author)

Abstract. Indonesia has the second-highest number of annual tuberculosis cases, and it has become a national problem. In 2016, the incidence of tuberculosis was approximately 39/100,000, with a mortality rate of 42/100,000 people. In addition, the current pandemic has made controlling infectious diseases more difficult. This study develops a machine learning model based on Principal Component Analysis-Artificial Neural Network (PCA-ANN) that can classify tuberculosis patient treatment failures based on 2,293 electronic medical record data. The socioeconomic data become the primary indicator because of the completeness of the data stored by the hospital. Some additional data, such as primary and secondary diagnoses coded using the International Classification of Diseases (ICD)-10, were included as medical indicators. The final status and the reason patients were discharged from the hospital are used to determine a failed or successful label. These data are processed in a dataset called the Socio-economic Tuberculosis Patient Indicator (STPI). This study uses the Cross Industry Standard Process for Data Mining (CRISP-DM) framework to develop the classification model. Two-dimensional PCA was used to reduce the variable dimensions in the dataset. Classification Model Development through the PCA-ANN architecture with the best results found in the accuracy of 97.21% for training and 96.51% for testing. The study results can be used to apply a classification model to address existing business needs in the context of tuberculosis control in Indonesia. Healthcare organizations/providers in Indonesia can use the steps and methods described in this study to develop their classification model according to the specified use case.

Keywords: Tuberculosis, Treatment Failure Classification, Artificial Neural Network, Principal Component Analysis, Crisp-Dm

1. Introduction

Tuberculosis is the most common infectious disease-related cause of death globally (World Health Organization, 2019). Including people living with Human *Immunodeficiency Virus* (HIV) infection (Gupta, Lucas, Fielding, & Lawn, 2015). The World Health Organization's (WHO) initiative, The Strategy to End TB, establishes ambitious targets for 2020-2035. By 2020, there will be a 20% reduction in incidence and a 35% reduction in tuberculosis fatalities, a 90% reduction in tuberculosis incidence and a 95% reduction in tuberculosis deaths by 2035, compared to 2015. (World Health Organization, 2015). A report has been published evaluating global progress against the target based on data from WHO (World Health Organization, 2019). Indonesia has the second-highest number of annual tuberculosis case,s and it has become a national problem. In 2017, the annual number of new tuberculosis cases rose to 25.40 per 1 million, with an 88% treatment success rate (World Health Organization, 2021). In 2016, the incidence of tuberculosis was approximately 39/100,000 people in Indonesia, with a mortality rate of 42/100,000 people (Asril, Soetikno, & Ascobat, 2019).

Diagnostic delay and treatment non-adherence contributed to this situation. COVID-19 may hinder Indonesia's efforts to eliminate tuberculosis (Caren, Iskandar, Pitaloka, Abdulah, & Suwantika, 2022). The SARS-CoV-2 virus causes Covid-19, PneumoViral infections cause pneumoniatuberculosis Mycobacterium Tuberculosis (MTB) bacteria cause tuberculosis and affects it. The current pandemic has made controlling infectious diseases more difficult. Although Indonesia is improving tuberculosis elimination, the current pandemic could hinder efforts. The pandemic would affect Indonesia's tuberculosis control (Caren et al., 2022; Pang, Liu, Du, Gao, & Li, 2020). The COVID-19 pandemic may exacerbate socioeconomic disparities, for example, due to relatively more significant increases in financial stress and suboptimal living conditions among individuals from low socioeconomic backgrounds (Stevens et al., 2022). Even though it's a big city, tuberculosis patients have common socioeconomic conditions in Jakarta. A total of 138 patients138 patients recovered, 85 died, and 66 stopped treatment from 2017-2020 in one private hospital. It means 138 successes and 151 failures occurred in this hospital. Many tuberculosis patients dropped out due to financial pressures, doctors said. Patients must pay for hospital transportation even though the government provides free tuberculosis drugs (Imam et al., 2021). Lost productivity and job losses are also important. Patients with limited education had trouble understanding instructions (Kalami, Woru, Tampubolon, & Handayani, 2020; Sahile, Yared, & Kaba, 2018).

Tuberculosis mainly affects those with a lower socioeconomic status in both high and low-income countries (Murray, 2004, 2015). Poverty, socioeconomic and gender inequalities, and living conditions primarily relevant to a tuberculosis risk (Janssens & Rieder, 2008). Poverty is defined by malnutrition, insufficient housing/living situations, and overpopulation (Carter et al., 2018). People from

lower socioeconomic groups are more likely to work in congested environments, be food insecure, be less aware of healthy habits, and have limited access to high-quality healthcare services. (Adler & Newman, 2002). They're also more likely to interact with those who have active tuberculosis. In diverse circumstances, people with low socioeconomic have a greater prevalence of tobacco use, alcohol usage, HIV, and diabetes mellitus (Agardh, Allebeck, Hallqvist, Moradi, & Sidorchuk, 2011; Gaskin et al., 2014; Karriker-Jaffe, Roberts, & Bond, 2013; Killewo, 2002). Although socioeconomic status has not consistently been demonstrated to be a factor in tuberculosis treatment failure, in underdeveloped nations, a low socioeconomic level can force patients to choose between competing priorities. Demands to direct the few resources available to address the needs of other family members, such as children or their beloved parents, are frequently among these priorities.

This study analyzes social and economic data from hospital medical records to improve treatment failure classification. When classifying certain classes, utilizing PCA will reduce the processing time by fifty percent. The general–purpose principal component analysis is useful for classifying data by reducing its dimensions (Kim & Kim, 2022). The ANN algorithm was chosen because it has the highest classification accuracy in studies (Hananti & Sari, 2021). The study included patients who recovered, improved, deteriorated, and died. The experiment tests the accuracy, precision, and recall of a socioeconomic classification for treatment failure in tuberculosis patients. The healthcare team will be informed about possible treatment failures based on the classification.

2. Literature Review

2.1. Electronic medical records

Electronic medical records (EMR) are commonly used to document patient conditions, including diagnostic information, procedures performed, and treatment outcomes. As a result, it has been acknowledged that EMR is a valuable resource for large-scale analysis. However, the EMR's characteristics of diversity, incompleteness, redundancy and privacy make it challenging to perform direct data mining and analysis (Sun et al., 2018; Mohammed and Fiaidhi 2022). The electronic record provides unmatched tools for increasing the precision and effectiveness of medication management. Medication lists can be automatically updated with each new prescription, and careful attention to comparing the drugs the patient is taking with the list of drugs helps to prevent forgetting to prescribe an essential medicine or prescribing a drug twice (Huber, Highland, Krishnamoorthi, & Tang, 2018). Vertical integration and care coordination are improved by sharing prescription lists among all healthcare professionals who treat the patient. The cost and insurance coverage of drugs are sometimes made available to the doctor at the point of care in EMR systems, enabling them to make more economically sensible prescription

selections. Prescribers and pharmacies are connected by highly advanced electronic prescribing systems, which eliminate the need for paper prescriptions and enhance the precision and efficiency of the prescribing and drug dispensing procedures (Janett & Yeracaris, 2020). Prediction models using EMR data typically perform more accurately than administrative data. However, this improvement is still only slight. Most of the studies under review did not discuss the clinical impact, did not calibrate the models, did not include socioeconomic features, or conduct rigorous diagnostic testing (Mahmoudi et al., 2020).

A study whose findings indicated that environmental risk factors statistically influenced the incidence of tuberculosis. These risk factors included patients with a history of contact, extensive house ventilation, inappropriate housing density, room humidity not up to standard, room temperature not following the standards, and the use of firewood for cooking. After multivariate analysis, the most significant risk factors for tuberculosis were patient contact, density, humidity, and temperature. The proficiency level for predicting the incidence of pulmonary tuberculosis of four risk factors were identified as 71.5% (Pratiwi, Pramono, & Junaedi, 2020). Their study investigated the socioeconomic predictors and distribution of positive tuberculosis in the Beijing region. The findings suggest that tuberculosis control measures should concentrate on these factors to allocate public health resources more effectively and reduce the incidence of tuberculosis (Mahara, Yang, Chen, Wang, & Guo, 2018).

While other studies focused on the relationship between the level of education, economic status, and smoking habits and the incidence of pulmonary tuberculosis in adults in Ketapang Regency, West Kalimantan. It was determined people with low levels of knowledge will increase their risk of developing tuberculosis by a factor of 1.857 relative to those with high levels of knowledge (Setiarni, Sutomo, & Hariyono, 2011). According to Purwanto's (2003) research, there is a correlation between the level of knowledge and the incidence of pulmonary tuberculosis. Then Imam et al. (2021) surveyed the potential influence of socioeconomic, income, and educational status on the adverse effects of drugs and their therapeutic episodes in patients with tuberculosis who received a combination of interventions. The result is the possibility of a high incidence of tuberculosis among patients with a secondary school education or less. In addition, the incidence is higher among patients with lower incomes than those with higher incomes. In addition, semi-skilled workers reported a higher incidence of adverse events than skilled workers. Moreover, the side effect was associated with increased proximity between patients from lower socioeconomic classes.

2.2. Cross-industry standard process for data mining (CRISP-DM)

According to Shearer (2000), The Cross Industry Standard Process for Data Mining (CRISP-DM) is an open-source methodology used by many data mining experts to develop analytical models. By Chapman et al (2000), This method was created as a

process based on a hierarchical model composed of various activities organized into four levels of abstraction: stages, general activities, specific actions, and procedures. CRISP-top DM's hierarchy, Phase, represents a collection of methodological activities. Each stage has a distinct objective and output. Each stage will consist of various general activities that will be carried out as part of the methodology's second level to accomplish the stated objectives and generate the outputs. Each of these activities can be carried out in various ways depending on the situation and circumstances, and as such, are classified as special activities – third level. The following stage will document each action, decision, and result, which will record the activities carried out using this methodology. The CRISP-DM has a well-defined life cycle that corresponds to data mining implementation.

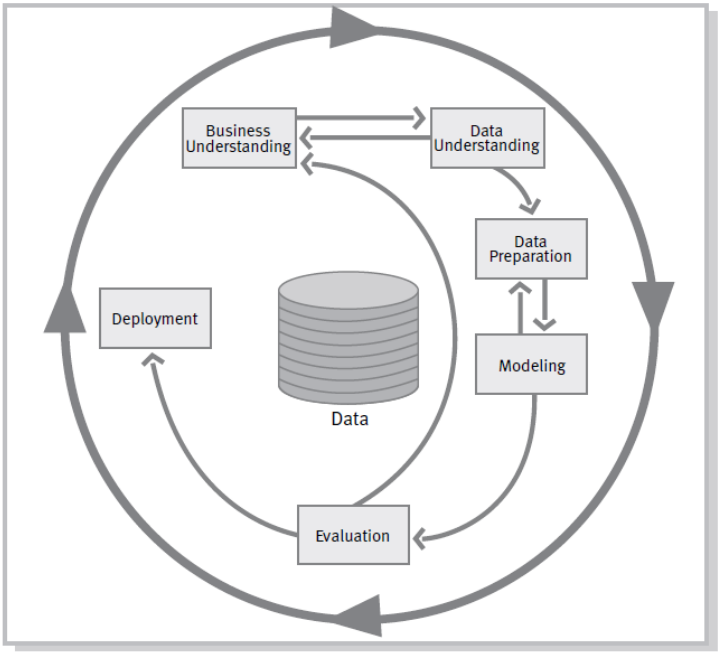

Fig. 1: CRISP-DM life cycle (Chapman et al., 2000)

Business Understanding - At this stage, data mining activities are carried out to ascertain the objectives and business requirements. Based on these two pieces of information, a definition of the problem to be solved is developed, as well as a plan for implementing the activities necessary to accomplish the stated goals.

Data Understanding - This stage collects and analyses initial data to improve understanding and recognition of the data, to accurately identify data quality, to gain initial insights, and to develop initial hypotheses based on detected data patterns.

Data Preparation - Various operations such as tabulation, transformation,

attribute/feature selection, and data cleaning/data cleaning are performed on the final dataset that will be used to create the model during this stage.

Modelling - Modelling stage activities employ a variety of modelling techniques and parameters with the goal of achieving the desired level of performance.

Evaluation - The Evaluation stage entails assessing the model's performance in order to ensure that all project implementation objectives, namely meeting identified business needs and resolving existing problems, have been met.

Deployment – At this last stage, a series of activities are carried out to ensure the value of the resulting benefits, as well as the implementation and use of the developed model.

2.3. Artificial neural network in healthcare

Artificial neural networks (ANNs) are designed to understand data, create new information based on learning, and utilize various variables. One of the activation functions is selected based on the dataset distribution. The weight values are automatically updated until the goal output values are attained in accordance with the learning principles (Manoj Kumar & Ananda, 2022). With the fusion of healthcare and ICT for real-time and quick results management, there has been a recent paradigm shift in digital healthcare services from treatment to prevention. However, despite the growing interest in using artificial intelligence (AI) based on healthcare data, medical institutions have been upgrading their diagnostic testing systems, arguably the most data-intensive of all methods used by medical institutions, to meet a variety of demands for health management (Oh, 2022).

Nachiappan & Devaraj (2021) conducted research on Remote Patient Diagnosis through IoT and Virtual Reality, Classification of Cloud Data Using ANN, utilizing the machine learning technique. The diagnostic procedure can be completed with a high level of dependability and efficacy using the exceptional medical environment devices. This study's primary objective is to investigate remote patient diagnosis via IoT and virtual reality, as well as the classification of cloud-based ANN implementations. Early identification of critical conditions, theories, approaches to infection identification, and methods for alert generation. Due to the limited time available to develop a bridge between the patient and the physician, the results of various diagnosis scenarios can also be evaluated in the case of measurements. k-Nearest Neighbor (kNN), Decision Tree (DT), Tabu Search (TS), Genetic Algorithm (GA), Naive Bayes (NB), Fuzzy Logic (FL), and Support Vector Machine are some of the classifiers that have been subdivided for analysis using the ANN technique, especially for diseases that include infection and heart problems (SVM). This study focuses on developing a treatment failure classification model for tuberculosis (TB) patients based on socioeconomic conditions using hospital-held medical record data. The model is constructed using the ANN and PCA models.

The combination of these two methods to classify tuberculosis treatment failure has never been used before, particularly for Indonesian patients.

2.4. PCA-ANN algorithm

Principal Component Analysis (PCA) is linear dimensionality reduction using Singular Value Decomposition of data to project it into a lower dimensional space and affected by scale. In the built model, StandardScaler is used to help standardize data set features to a unit scale (mean = 0 and variance = 1), which is a requirement for the optimal performance of many machine learning algorithms (Granato, Santos, Escher, Ferreira, & Maggio, 2018). The PCA is used to extract feature vectors by reducing the dimensions, and ANN is used in the classification algorithm (Bayar, Terzi, & Ozgonenel, 2019). It is a technique for reducing the dimensionality of a data set, improving interpretation while minimizing the loss of information. In addition, PCA will create new variables, and principal components are reduced to solving eigenvalue/eigenvector problems, and new variables are determined by existing data sets, not a priori, thus making PCA an adaptive data analysis technique (Jolliffe & Cadima, 2016).

The use of PCA is expected to improve the accuracy performance of the classification model. In other study, using examples of early questionnaire screening, machine learning can offer a pattern for the classification of cancer types, focusing on lung, liver, upper, upper, lower, and breast cancer. In this study, 3411 respondents were screened. Cancer can be predicted by incorporating 28 attributes into models created using principal component analysis (PCA) and artificial neural network (ANN) technology (Qi et al., 2022). In another study, PCA-ANN was used to measure a person's body temperature. The proposed multimodality sensor (MMS) system consists of a thermopile infrared sensor, a proximity sensor, a light intensity sensor, and a web camera. Furthermore, a principal component analysis (PCA) approach incorporating artificial neural networks is used for body temperature regression. To reduce measurement bias, digital axillary temperature data is used as baseline body temperature for machine learning training and validation. The results show that the PCA-ANN temperature estimation model outperforms other prediction models with the best average (Rinanto & Kuo, 2021). PCA-ANN has also been used to identify diabetes in other studies. For example, principal Component Analysis (PCA) and Multilayer Perceptron Artificial Neural Networks are used to create a diabetes detection system. The focus of the investigation is to integrate PCA's data source and transformation capabilities into the MLPNN framework. A confusion matrix-based analysis was used to examine the impact of source information fusion and PCA. A maximum accuracy of 76.5% was recorded for the source feature in the standard UCI diabetes dataset analysis, and an accuracy of 85.2% was recorded for the 6-level PCA feature. Fusion has the highest

success rate at 87.8%. Compared to the source features and PCA used alone, the relative accuracy increases by 15% and 3%, respectively (Sangle, Kachare, & Sonawane, 2019).

3. Methodology

3.1. Research Design

The stages of development of the classification model are adjusted to the type of activity to be carried out as part of the scope of the study, namely the creation of a classification model for treatment failure of tuberculosis patients based on socioeconomic conditions based on Artificial Neural Network (ANN), and its application using the CRISP-DM methodology. The steps carried out are as follows:

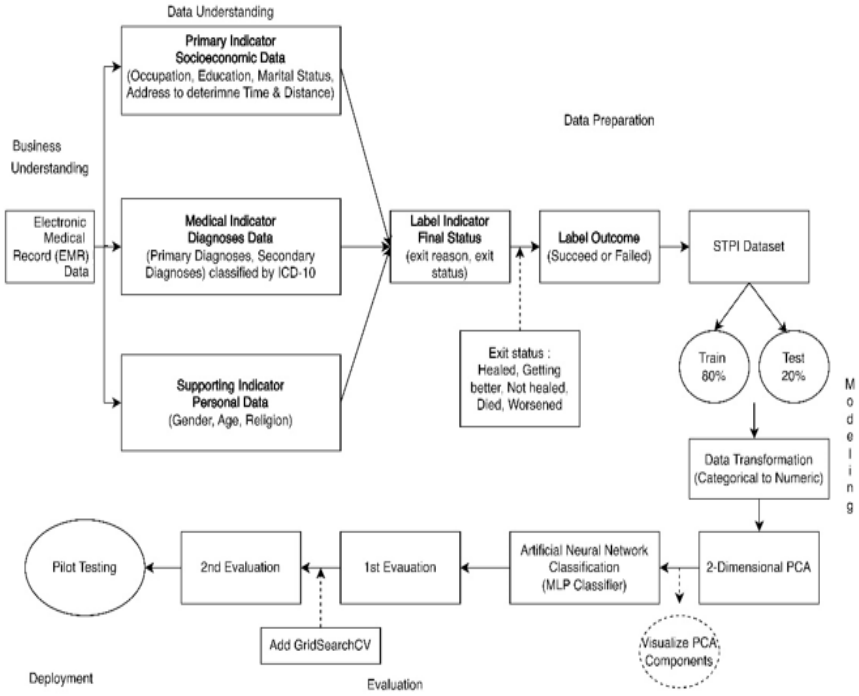

Fig. 2: Research design

This study begins with data collection through the hospital's electronic medical record (EMR) data. In the patient's medical record data, there is much information that can be used for the construction of a classification model. In this study, the CRISP-DM framework is used. It is necessary to know the reasons that influence the failure of tuberculosis treatment and the information related to this occurrence. Understanding this electronic medical record data is an understanding of the business processes carried out in medical activities. In the next stage, we divide the medical record data into three indicators, including primary, medical, and support.

The primary indicators are socioeconomic data, and medical data are primary and secondary diagnoses of tuberculosis patients, which in the EMR data have been classified using ICD-10. This stage is called Data Understanding, after we process these data into a dataset with indicator labels that are supported by two final patient statuses, including the reason they were discharged from the hospital and the final status before they were discharged. This indicator label determines Succeed or Failed.

The whole process, from business understanding to data preparation, produces a dataset called the Socioeconomic Tuberculosis Patient Indicator (STPI). First, the STPI dataset is split into two parts, 80% as a training set and 20% as a test set. Both datasets contain categorical data, which is transformed into numeric. PCA is used at this modeling stage. Next, the multidimensional data is simplified into only two dimensions called Principal Components. After that, these two datasets were then trained and tested using a classification algorithm, MLP Classifier. The results of the training and testing of these two datasets are evaluated in the first evaluation stage using the Confusion Matrix to see the accuracy value. Then, the addition of GridSearchCV is done to increase the level of accuracy and to experiment with models that occur overfitting or underfitting. After that, a second evaluation stage was carried out to see the accuracy results after the addition of GridSearchCV. Finally, the experiments' models are compared so that conclusions can be drawn about which model has the highest accuracy value. This model is called the Best Model and can be used for the deployment stage, exactly at the pilot testing stage at the hospital.

3.2. Data collection method

Data collection is carried out by involving retrospective data on tuberculosis patients who have been and are undergoing treatment for a certain period in the last three years, mainly from 2019 until. 2021. This data will be used in developing a classification model, which takes up to 3 months, **from February 2022 to April 2022**. The study was conducted at the Central General Hospital in East Jakarta. This location was chosen because it has many tuberculosis patients each year, making it possible to build a classification model.

3.3. Sample and Setting

The population for this study is tuberculosis patients who have received or are currently receiving treatment at the Hospital. This study employs a purposive side technique, selecting a research sample with specific considerations to ensure that the data obtained later are representative. Sample management is accomplished through the Python programming language, the machine learning software Anaconda Navigator and Jupyter Notebook, and the number processing software Microsoft Excel. Sample calculations for data mining using ANN sample size should be at least 50-500 times the number of features (Alwosheel, van

Cranenburgh, & Chorus, 2018). Because in this study there were five main features, we took the number of calculations for the sample size in this study which was 450 times the number of main features which resulted in 2,250 data.

3.4. Ethical Clearance

All data taken from the patient's medical record will be kept confidential and used for this study only, and we received approval to pass the ethical review from the Health Research Ethics Committee of Central General Hospital. The ethical clearance number for this research is 12/KEPK-RSUPP/02/2022, and permission to conduct research with the letter DL.01.01/IX.2/2059/2022.

4. Results and Discussion

4.1. Business Understanding

The classification model should classify patient treatment failures based on electronic medical record data. The hospital's current medical record form includes socioeconomic data. No tuberculosis classification model meets this technical requirement. The study propose a 95% accurate ANN classification model for large datasets (Wang, Lin, & Dang, 2020). The classification model will be built using tuberculosis patient medical records from the last three years. No tuberculosis patient dataset exists to build the model. In this study, it is necessary to create a dataset at Central General Hospital to represent socio-economic indicators such as education level, occupation, marital status, travel time to the hospital, and some medical data as supporting factors.

4.2. Data understanding

In the medical record data provided by the hospital, there are several data that can be used as a reference in terms of the socioeconomic conditions of tuberculosis patients, including:

- **Occupation Data** - Retirees, Housekeeping, Private Employees, Civil Servants, Others, Unemployed, Students, Nuns, Entrepreneurs, Traders, Freelancer, Drivers, State-Owned Company Employees, Farm Workers / Plantation, Doctors, Nurses, Teachers, Indonesian National Armed Forces, Lecturers, Artists, Farmers / Planters, Transportation, Mechanics, Parliamentarian, and Journalists.
- **Education Level Data** - Unschooled, Kindergarten, Primary School/equivalent, Junior High School/equivalent, Senior High School/equivalent, Associate's I/II, Academy/Associate's III/ Bachelor, Master, Doctoral.
- **Patient's address** - given in the form of complete address. And identified by villages address to see the distance and travel time to the hospital.
- **Marital status data** - Single, Married, Divorced.

Supporting factors in developing this classification model were recommended by a team of doctors from Central General Hospital when submitting research ethics. Some of the recommended supporting data include:

- Gender, because there are physiological and psychological differences between men and women. This affects treatment success.
- Age is very influential, especially in terms of dosing, monitoring, and treatment procedures.
- Religion is a consideration because tuberculosis drugs must be taken regularly at the same time every day. It is a consideration for some religions that have special drug and food consumption circumstances.

4.3. Data preparation

Collecting data is the first step in preparing the Socioeconomic Tuberculosis Patient Indicators (STPI) dataset. This study used sensitive medical record data before data collection, and a review of research ethics must be conducted. An application must pass several review processes and an ethics committee trial to be approved. This research was ethically approved for data collection. All data from the patient's medical record will be kept confidential and used for this study only. Medical record spreadsheet exported from the hospital IT team. Each year's data is on a separate worksheet. This study uses data from 2019 to 2021 containing up to 2,293 rows of data, per hospital data retention policies. Medical record data is transformed into a dataset with socioeconomic indicators and other variables. The STPI dataset must be pre-prepared for this study. Medical record data is exported from the hospital information system as**.xlsx** spreadsheets. Some data need adjusting:

- Because of too many different values, secondary diagnosis data are grouped as present or absent.
- Data are grouped by 10km or less. 10km is based on 2014 Village Potential (PODES) data, which provides information about distance and access to health facilities (Maisya & Lestari, 2019);
- The average travel time per 10km in Jakarta is 30 minutes.

4.4. Modeling

The split train-test procedure estimates machine learning algorithm performance on unlabeled data, 80:20 training-to-testing data ratio with random state parameter 44. The used dataset is categorical. Sklearn expects categorical data to be numbers. If data is numerical, training can begin (Manaswi & John, 2018). The encoding label specifies string-to-number conversion. "Label encoding" converts text labels to numbers. With default parameters, MLPClassifier (ANN) trains the model. It is a Neural Networks-related Multi-layer Perceptron classifier that uses a Neural Network for classification, unlike SVMs or Naive Bayes (Zhang et al., 2018). At this stage, a classification model has been created to predict test results. The MLPClassifier is used in classification prediction problems with labeled input, and

only the random state has a value in the built model for reproducing results. The model predicts test data. Keras is a high-level deep-learning API written in Python. It uses a simple API to build a model, train it, and then use it for predictions (Manaswi, 2018; Moolayil, 2019).

The study added PCA to condenses data, improving interpretation and minimize information loss. PCA generates new variables and reduces the main component to solving eigenvalues/eigenvectors, making it an adaptive data analysis technique (Jolliffe & Cadima, 2016). We created 2 PCA dimensions into our model. GridSearchCV optimization used to increases testing accuracy and prevents overfitting. Due to imbalanced data, f1 weighted scoring is needed. Averages all F1 scores across all classes while taking each class's support into account. This optimization model ensures the best model is obtained and post-optimization model performance is evaluated. The optimization parameters are:

- alpha: [0.0001, 0.0003, 0.001, 0.003],
- hidden_layer_sizes: [(16,), (32, 16), (32, 16, 8), (100,), (105,)],
- learning_rate_init: [0.001, 0.005, 0.01],
- activation: ["relu", "logistic", "tanh"]

While there are other ways to speed up machine learning algorithms, one less commonly known method is to use PCA. The graph above shows that the classes seem well separated from each other. The explained variance tells you how much information (variance) can be attributed to each of the principal components. This is important as we can convert multidimensional space to 2-dimensional space. Using the attribute explained_variance_ratio can see that the first principal component contains 75.77% of the variance, and the second principal component contains 24.23%. Together, the two components still retain 100% of the information.

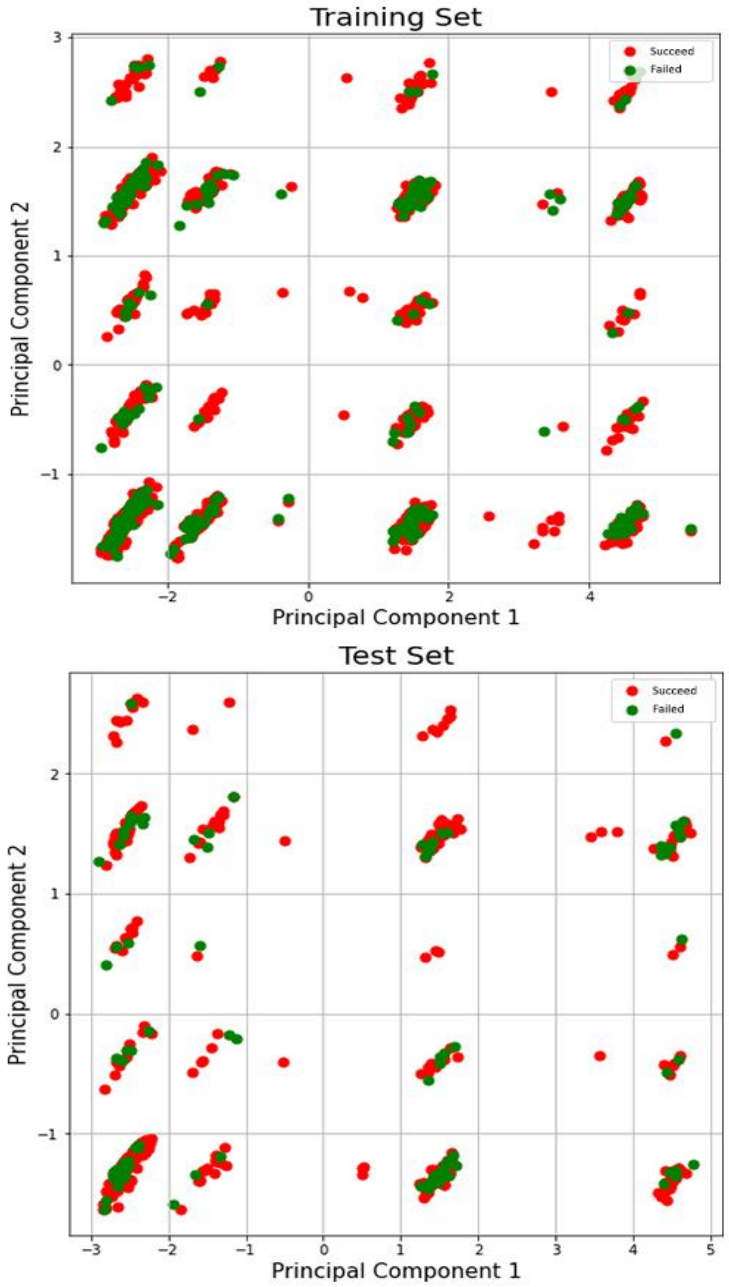

Fig. 3: Principal components plot in training and test set

4.5. Evaluation

The performance of each model in the experiments conducted. It is a report on performance comparisons derived from the confusion matrix shown in Table 1.

Table 1: Confusion Matrix results of all experiments

Experiment	*True Positive* (TP)	*True Negative* (TN)	*False Positive* (FP)	*False Negative* (FN)
ANN	232	15	144	68
ANN-PCA	375	68	1	15
ANN + GridSearchCV	5	47	371	36
ANN-PCA + GridSearchCV	368	75	8	8

Table 1. shows that the best classification results are ANN-PCA + GridSearchCV experiment. In this experiment, the data that were correctly classified for the success label (TP) was 368 (97.87%), reducing 8 data from before optimization, and the failed label (TN) was classified correctly was 75 (90.36%) up 8 data from before optimization. This experiment gives the largest number of TP and TN from other experiments.

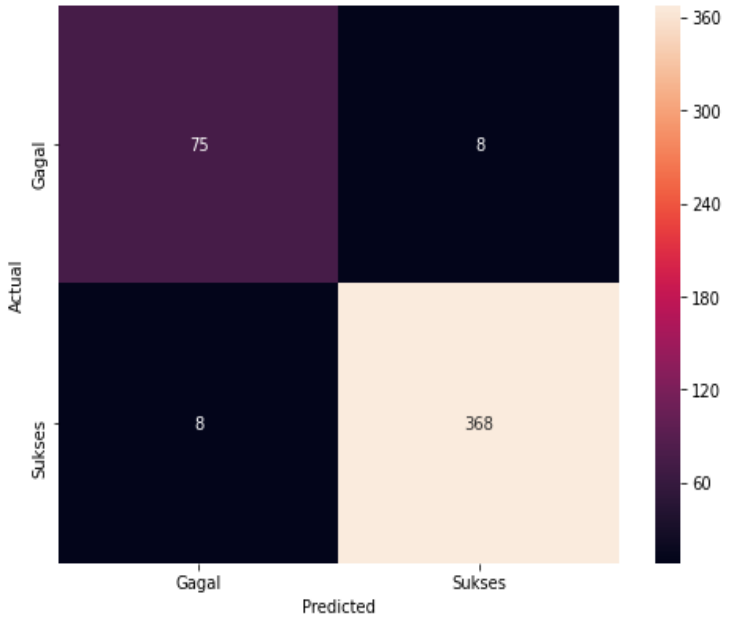

Fig. 4: Confusion matrix of experiment ANN-PCA + GridSearchCV

In the confusion matrix ANN-PCA experiment after optimization with GridSearchCV, which we can see in Figure 3, it states that on the Failed label, 75 data are correctly predicted (TN), and only 8 data are incorrect (FN). Meanwhile, the data for the Succeed label is predicted to be correct as much as 368 data (TP), and 8 data is predicted to be incorrect (TN). The experiment is the best because it can classify succeed and failed treatment of tuberculosis patients according to their socio-economic conditions. The overall level of accuracy of post-optimization experiments can be seen in table 2.

Table 2: Model accuracy rate

Experiment	Training	Testing	State
ANN	99.7%	53.8%	*Overfitting*
ANN-PCA	96.7%	96.5%	Better
ANN + GridSearchCV	100%	11.32%	*Overfitting - decreased*
ANN-PCA + GridSearchCV	97.21%	96.51%	Best - increased

From the table above, it can be concluded ANN-PCA + GridSearchCV experiment is a classification model with the best level of accuracy in terms of testing and tends to increase in post-optimization training. Moreover, the model in this experiment does not experience overfitting as in the other two experiments. For more details regarding the performance of the classification model for each experiment, we can see the report in Table 3.

Table 3: Classification results report

Experiment	Precision		Recall		F1-Score	
	Failure Label	Success Label	Failure Label	Success Label	Failure Label	Success Label
ANN	9%	77%	18%	62%	12%	69%
ANN-PCA	99%	96%	82%	100%	89%	98%
ANN + GridSearchCV	11%	12%	57%	1%	19%	2%
ANN-PCA + GridSearchCV	24%	98%	98%	31%	38%	47%

From several evaluation results that have been reported in this section, it can be concluded that ANN-PCA experiment creates the better classification modelc, and became the best after optimization using GridSearchCV. The model in ANN-PCA + GridSearchCV experiment can be said to be the best model. To ensure that there is a consistent performance in this model, the pre- and post-optimization performance can be described in table 4.

Table 4: Best model performance comparison

Performance	Accuracy		Precision		Recall		F1-Score	
	Training	Testing	Failure Label	Success Label	Failure Label	Success Label	Failure Label	Success Label
Before Optimization	96.7%	96.5%	99%	96%	82%	100%	89%	98%
Post Optimization	97.21%	96.51%	90%	98%	90%	98%	90%	98%

From the comparison results, it can be concluded that the ANN-PCA experiment had good performance before GridSearchCV optimization. However, some aspects of performance, such as accuracy, improved after optimization. This comparison can be used to conclude that the model developed in this experiment remains consistent and becomes the best before and after optimization, even though this increase is not very significant.

4.6. Deployment

During a period of limited/pilot deployment, the application was evaluated on several new patients to determine the system's readiness. Additionally, the system's effects on the privacy and security of associated data must be considered. The model was tested on several new patients. The best model ANN-PCA algorithm was then used to classify the newly acquired data. This model will be installed on the current information system of the hospital. To accomplish this, a study must be conducted with the IT team so that its application can be utilized in future research on the development of information systems.

Theoretically, this study illustrates how influential socio-economic conditions are on the treatment failure of tuberculosis patients. When talking about treatment, it cannot depend solely on medical data. Some data supporting the success of treatment must be considered and utilized in such a way that it can be used as helpful information in the success of the process. And practically, to apply the tuberculosis treatment failure classification model, collecting and using datasets from patients who have recovered or indicated treatment failure is recommended to ensure adequate model accuracy. Activities to collect and prepare medical record data must protect the security and confidentiality of the personal data collected, so there is a need for an ethical permit from the hospital/organization that stores the patient data. Some procedures are needed to adopt the model to the current information system in the hospital/organization. Further research can focus on examining the application of the tuberculosis treatment failure classification model for various use cases, such as synchronization with hospital medical record data, monitoring patient medication intake, and identifying adherence during treatment.

5. Conclusion

This study developed the Socioeconomic Tuberculosis Patient Indicators (STPI) dataset for Indonesia, along with an explanation of the development process. This dataset includes 2,293 tuberculosis patient records from Central General Hospital. The records include education, occupation, marital status, age, gender, religion, travel time to the hospital, and therapy-related health conditions. Classification models can be built using the best ANN architecture in the experiment with a

training accuracy of 97.21% and a testing accuracy of 96.50% using PCA and GridSearchCV. The classification model's predictive/inference ability was also evaluated. Study findings can be used to develop a classification model for tuberculosis control in Indonesia. Indonesian healthcare organizations/providers can use the study's steps, methods, and downloadable source code to create their classification model for the specified use case. This study's findings can be used to classify tuberculosis treatment failure based on socioeconomic status.

This study only focuses on developing a basic classification model. Thus, an entirely produced application has yet to be used by health workers and patients. Then, the dataset used in this study is electronic medical records (EMR), thus tends to be challenging to get the same good results if applied to health facilities that were not implementing the EMR. For future works, developing an application that can adapt the classification model in this study is necessary to use both the EMR and not the EMR dataset.

References

Adler, N. E., & Newman, K. (2002). Socioeconomic Disparities in Health: Pathways and Policies. Health Affairs (Project Hope). 21(2), 60–76. https://doi.org/10.1377/hlthaff.21.2.60

Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., & Sidorchuk, A. (2011). Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. International Journal of Epidemiology, 40(3), 804–818. https://doi.org/10.1093/ije/dyr029

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your Dataset Big Enough? Sample Size Requirements when using Artificial Neural Networks for Discrete Choice Analysis. Journal of Choice Modelling, 28, 167–182. https://doi.org/https://doi.org/10.1016/j.jocm.2018.07.002

Asril, I., Soetikno, V., & Ascobat, P. (2019). Associations between the Adverse Drug Reactions and the Tuberculosis Treatment Dropout Rates at the Cempaka Putih Islamic Hospital in Jakarta, Indonesia. Journal of Natural Science, Biology and Medicine, 10(3), 29–33

Bayar, H., Terzi, U. K., & Ozgonenel, O. (2019). PCA-ANN based Algorithm for the Determination of Asymmetrical Network Failures of Network-Connected Induction Generators. Tehnički Vjesnik, 26(4), 953–959

Caren, G. J., Iskandar, D., Pitaloka, D. A. E., Abdulah, R., & Suwantika, A. A. (2022). COVID-19 Pandemic Disruption on the Management of Tuberculosis Treatment in Indonesia. Journal of Multidisciplinary Healthcare, 15, 175–183. https://doi.org/10.2147/JMDH.S341130

Carter, D. J., Glaziou, P., Lönnroth, K., Siroka, A., Floyd, K., Weil, D., Boccia, D. (2018). The Impact of Social Protection and Poverty Elimination on Global Tuberculosis Incidence: A Statistical Modelling Analysis of Sustainable Development Goal 1. The Lancet. Global Health, 6(5), e514–e522. https://doi.org/10.1016/S2214-109X(18)30195-5

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM -Cross-Industry Standard Process for Data Mining-1.0 Step-by-step data mining guide. CRISP-DM Consortium.

Gaskin, D. J., Thorpe, R. J. J., McGinty, E. E., Bower, K., Rohde, C., Young, J. H., Dubay, L. (2014). Disparities in Diabetes: the Nexus of race, Poverty, and Place. American Journal of Public Health, 104(11), 2147–2155. https://doi.org/10.2105/AJPH.2013.301420

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) for Multivariate Association between Bioactive Compounds And Functional Properties in Foods: A Critical Perspective. Trends in Food Science & Technology, 72, 83–90

Gupta, R. K., Lucas, S. B., Fielding, K. L., & Lawn, S. D. (2015). Prevalence of Tuberculosis in Post-Mortem Studies of HIV-Infected Adults and Children in Resource-Limited Settings: A Systematic Review and Meta-Analysis. AIDS (London, England), 29(15), 1987

Hananti, H., & Sari, K. (2021). Perbandingan Metode Support Vector Machine (SVM) dan Artificial Neural Network (ANN) pada Klasifikasi Gizi Balita. Seminar Nasional Official Statistics, 2021(1), 1036–1043

Huber, M. T., Highland, J. D., Krishnamoorthi, V. R., & Tang, J. W.-Y. (2018). Utilizing the Electronic Health Record to Improve Advance Care Planning: A Systematic Review. American Journal of Hospice and Palliative Medicine®, 35(3), 532–541

Imam, F., Sharma, M., Al-Harbi, N. O., Khan, M. R., Qamar, W., Iqbal, M., Anwar, M. K. (2021). The Possible Impact of Socioeconomic, Income, and Educational Status on Adverse Effects of Drug and their Therapeutic Episodes in Patients Targeted with a Combination of Tuberculosis Interventions. Saudi Journal of Biological Sciences, 28(4), 2041–2048

Janett, R. S., & Yeracaris, P. P. (2020). Electronic Medical Records in the American Health System: challenges and lessons learned. Ciencia & Saude Coletiva, 25, 1293–1304.

Janssens, J.-P., & Rieder, H. L. (2008, November). An Ecological Analysis of Incidence of Tuberculosis and Per Capita Gross Domestic Product. The European Respiratory Journal. 32, 1415–1416. https://doi.org/10.1183/09031936.00078708

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202

Kalami, M. W., Woru, H. Z., Tampubolon, E., & Handayani, D. (2020). Association of Socioeconomic Factors and Medication Compliance Among Drug Resistant Tuberculosis Patients in West Papua, Indonesia. In D36. International Perspectives on Pulmonary And Critical Care Medicine (pp. A6560–A6560). American Thoracic Society

Karriker-Jaffe, K. J., Roberts, S. C. M., & Bond, J. (2013). Income inequality, alcohol use, and alcohol-related problems. American Journal of Public Health, 103(4), 649–656. https://doi.org/10.2105/AJPH.2012.300882

Killewo, J. (2002, December). Poverty, TB, and HIV Infection: A Vicious Cycle. Journal of Health, Population, and Nutrition, 20, 281–284. Bangladesh

Kim, S.-C., & Kim, Y.-H. (2022). Efficient Classification of Human Activity using PCA and Deep Learning LSTM with WiFi CSI. 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 329–332. https://doi.org/10.1109/ICAIIC54071.2022.9722627

Mahara, G., Yang, K., Chen, S., Wang, W., & Guo, X. (2018). Socio-economic Predictors and Distribution of Tuberculosis Incidence in Beijing, China: A Study using a Combination of Spatial Statistics and GIS Technology. Medical Sciences, *6*(2), 26

Mahmoudi, E., Kamdar, N., Kim, N., Gonzales, G., Singh, K., & Waljee, A. K. (2020). Use of Electronic Medical Records In Development and Validation of Risk Prediction Models of Hospital Readmission: Systematic Review. Bmj, 369

Maisya, I. B., & Lestari, H. (2019). Jaminan Kesehatan sebagai solusi mengatasi hambatan akses dan biaya dalam pemanfaatan pelayanan kesehatan ibu di Indonesia. Badan Penelitian Dan Pengembangan Kesehatan

Manaswi, N. K. (2018). Understanding and working with Keras. In Deep Learning with Applications Using Python (pp. 31–43). Springer.

Manaswi, N. K., & John, S. (2018). Deep Learning with Applications using Python. Springer.

Manoj Kumar, D. P., & Ananda, B. J. (2022). Neural Network-based Game Theory Approach for Personalized Privacy Preservation in Data Publishing. Journal of System and Management Sciences, 12(1), 498–520. https://doi.org/10.33168/JSMS.2022.0132

Mohammed, S. & Fiaidhi, J. Extending the Power of Problem Oriented Medical Record with Disease Association Discovery: The Case Study of Empowering

QL4POMR with OpenTargets. International Journal of Hybrid Information Technology, 2(1), 1-12. https://doi.org/10.21742/IJHIT.2022.2.1.01

Moolayil, J. (2019). An Introduction to Deep Learning and Keras. In Learn Keras for Deep Neural Networks. 1–16. Springer.

Murray, J. F. (2004). A century of tuberculosis. American Journal of Respiratory and Critical Care Medicine, 169(11), 1181–1186. https://doi.org/10.1164/rccm.200402-140OE

Murray, J. F. (2015). Tuberculosis and World War I. American Journal of Respiratory and Critical Care Medicine, 192(4), 411–414. https://doi.org/10.1164/rccm.201501-0135OE

Nachiappan, V. A., & Devaraj, R. (2021). Remote Diagnosis of the Patient through IOT and Virtual Reality, Classification of the Cloud Data using ANN. REVISTA GEINTEC-Gestao Inovacao E Tecnologias, 11(1), 6025–6034

Oh, J. W. (2022). A Study on Digital Healthcare Service in Big Data Environment: Focusing on Diagnosis of Hyperlipidemia Based on Diagnostic Testing. Journal of System and Management Sciences, 12(3), 345–360. https://doi.org/10.33168/JSMS.2022.0317

Pang, Y., Liu, Y., Du, J., Gao, J., & Li, L. (2020, May). Impact of COVID-19 on tuberculosis control in China. The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease. 24, 545–547. https://doi.org/10.5588/ijtld.20.0127

Pratiwi, R. D., Pramono, D., & Junaedi, J. (2020). Socio-Economic and Environmental Risk Factors of Tuberculosis in Wonosobo, Central Java, Indonesia. KEMAS: Jurnal Kesehatan Masyarakat, 16(1), 61–70

Qi, H., Xie, S., Chen, Y., Wang, C., Wang, T., Sun, B., & Sun, M. (2022). Prediction Methods of Common Cancers in China Using PCA-ANN and DBN-ELM-BP. IEEE Access

Rinanto, N., & Kuo, C.-H. (2021). PCA-ANN Contactless Multimodality Sensors for Body Temperature Estimation. IEEE Transactions on Instrumentation and Measurement, *70*, 1–16

Sahile, Z., Yared, A., & Kaba, M. (2018). Patients' experiences and perceptions on associates of TB treatment adherence: A qualitative study on DOTS service in public health centers in Addis Ababa, Ethiopia. BMC Public Health, 18(1), 1–12. https://doi.org/10.1186/s12889-018-5404-y

Sangle, S., Kachare, P., & Sonawane, J. (2019). PCA fusion for ANN-based diabetes diagnostic. In Computing, Communication and Signal Processing (pp. 583–590). Springer

Setiarni, S. M., Sutomo, A. H., & Hariyono, W. (2011). Hubungan antara tingkat pengetahuan, status ekonomi dan kebiasaan merokok dengan kejadian tuberkulosis paru pada orang dewasa di wilayah kerja puskesmas tuan-tuan kabupaten ketapang kalimantan barat. Kes Mas: Jurnal Fakultas Kesehatan Masyarakat Universitas Ahmad Daulan, 5(3), 25008

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22

Stevens, G. W. J. M., Buyukcan-Tetik, A., Maes, M., Weinberg, D., Vermeulen, S., Visser, K., & Finkenauer, C. (2022). Examining socioeconomic disparities in changes in adolescent mental health before and during different phases of the coronavirus disease 2019 pandemic. Stress and Health, n/a(n/a). https://doi.org/https://doi.org/10.1002/smi.3179

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: A review. Journal of Healthcare Engineering, 2018

Tobacco and poverty : a vicious circle. (n.d.). Retrieved March 13, 2020, from https://apps.who.int/iris/handle/10665/68704

Wang, X., Lin, X., & Dang, X. (2020). Supervised learning in spiking neural networks: A review of algorithms and evaluations. Neural Networks, 125, 258–280.

WHO. (2021). Global Tuberculosis Report 2021 (2021st ed.; World Health Organization, Ed.). World Health Organization

World Health Organization. (2015). The end TB strategy. World Health Organization

World Health Organization. (2019). Global tuberculosis report 2019. Retrieved from https://apps.who.int/iris/handle/10665/329368

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., & Atkinson, P. M. (2018). A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 140, 133–144