Inteligencia Artificial 24(68), 104-122 doi: 10.4114/intartif.vol24iss68pp104-122



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

An Accurate Integrated System to detect Pulmonary and Extra Pulmonary Tuberculosis using Machine Learning Algorithms

Rupinder Kaur^[1], Anurag Sharma^[2]

^[1]Research Scholar, GNA University, Phagwara, Punjab, India rupinder8182@gmail.com

^[2]Professor cum Head, Faculty of Engineering Design and Automation, GNA University, Phagwara, Punjab, India anurag.sharma@gnauniversity.edu.in

Abstract Several studies have been reported the use of machine learning algorithms in the detection of Tuberculosis, but studies that discuss the detection of both types of TB, i.e., Pulmonary and Extra Pulmonary Tuberculosis, using machine learning algorithms are lacking. Therefore, an integrated system based on machine learning models has been proposed in this paper to assist doctors and radiologists in interpreting patients' data to detect of PTB and EPTB. Three basic machine learning algorithms, Decision Tree, Naïve Bayes, SVM, have been used to predict and compare their performance. The clinical data and the image data are used as input to the models and these datasets have been collected from various hospitals of Jalandhar, Punjab, India. The dataset used to train the model comprises 200 patients' data containing 90 PTB patients, 67 EPTB patients, and 43 patients having NO TB. The validation dataset contains 49 patients, which exhibited the best accuracy of 95% for classifying PTB and EPTB using Decision Tree, a machine learning algorithm.

Keywords: Machine Learning, Pulmonary Tuberculosis, Extra Pulmonary Tuberculosis, Chest X-Ray

1 Introduction

Tuberculosis is a global health problem that affects about a quarter of the world's population (WHO, 2020). In 2019, 10 million people became ill with tuberculosis, of which 56% were men, 32% were women, and 12% were children who developed TB. Furthermore, there are 30 high TB burden countries reported by WHO TB Report 2020, from which eight countries are considered as highest TB affected countries, with India leading the count, i.e., 26% followed by Indonesia (8.5%), China (8.4%), the Philippines (6%), Pakistan (5.7%), Nigeria (4.4%), Bangladesh (3.6%) and South Africa (3.6%). While most patients are infected with TB in the lungs, referred to as PTB, some individuals might have TB infection in other parts of the body, also referred to as EPTB. Hence, the diagnosis of TB is relatively more complex, unlike other infectious diseases. Various conventional methods are used to diagnose TB, but WHO's Global Tuberculosis Report says that intensified efforts are needed to improve access to diagnosis and treatment, especially in countries that accounted for more than half of the global gap and India is one of them (WHO,

2020). Therefore, machine learning algorithms are used in this research to detect both types of TB as they can learn from available features of data and then make predictions on blind data.

Fewer researches explored that studies related to EPTB are less in comparison with PTB. Among the collected datasets of TB patients, about 30% to 50% cases are of EPTB (Prakasha *et al.*, 2013), (Shrivastava *et al.*, 2015), (Velingker, Lawande and Dcosta, 2018). Pune Municipal Corporation recorded that the number of EPTB cases increased from 1,334 to 2,600 in three years (from 2017 to 2019), which means the cases almost doubled (TOI, 2020). The EPTB is difficult to diagnose due to its pauci-bacillary nature. The EPTB is mainly diagnosed using tissue biopsy, Ziehl-Neelsen stain, Fluorescent microscopy, Interferon-γ release assay. These methods need proper laboratory facilities, which are lacking in resource-limited settings. Also, to secure the samples of patients, there is a need for invasive procedures which sometimes become inaccessible (Purohit and Mustafa, 2015). Similarly, for PTB diagnosis, various methods such as sputum smear, culture and biopsies are used, which are very time consuming. Therefore, in this research, EPTB and PTB are diagnosed using machine learning algorithms to make early diagnosis possible at low cost and good precision.

In this work, a system is proposed that can diagnose both PTB and EPTB. For this, various meetings were done with doctors and experts to finalize input parameters. Then data is collected from several hospitals of Jalandhar. This dataset includes clinical data, hematological data, and site-related data. The PTB parameters include the basic parameters that the TB patient has whereas the EPTB parameters include the basic as well as the site-related parameters. CXRs (X-Rays) are also collected for TB patients as it the primary tool used for the detection of TB as it is proven to be efficient, cost-effective and it is having high sensitivity (Guo, Passi and Jain, 2020), (Sathitratanacheewin, Sunanta and Pongpirul, 2020b). For PTB, X-rays of lungs are collected whereas, for EPTB, X-Rays of that particular site are collected where the TB is supposed to be. After collecting the dataset, preprocessing is done which is done by removing missing data and doing some transformations, similarly for images the normalization is done. In the next step, three basic machine learning algorithms, i.e., Decision Tree, Naïve Bayes, and Support Vector Machine (SVM) are chosen to detect tuberculosis, classifying the dataset into three categories: normal, PTB and EPTB. In the last, performance is evaluated on the basis of five performance evaluation parameters i.e. AUC value, Accuracy, Precision, Recall and F1 Score.

2 Literature Survey

In this section, a number of TB diagnostic methods based on machine learning and deep learning are explored. To prevent the growth of the infection, early detection of TB is a must. However, the lack of skilled radiologists becomes a major hindrance in diagnosing TB effectively. Computer technology can provide very significant results compared to human diagnostics and make it possible to make the system with good accuracy and low cost. Table 1 shows the comparison of Tuberculosis detection methods.

Researchers present a TB diagnostic solution using the Bayesian Convolution Neural Network rather than a simple Convolution Neural Network. They implement it on the Google Colab platform using two TB datasets, namely Montgomery and Shenzhen, and have achieved 96.42% and 86.46% accuracy on both datasets, respectively (Ul Abideen *et al.*, 2020).

Researchers proposed the system to detect Tuberculosis reliably using Chest X-rays with segmentation, visualization, and deep learning classification techniques. Nine different deep learning models (ResNet18, ResNet50, ResNet101, ChexNet, InceptionV3, Vgg19, DenseNet201, SqueezeNet, and MobileNet) were used for transfer learning that will classify TB and non TB cases. U-net models were used for segmentation and X-ray images were used for classification. An experiment was also done using segmented lung images. The system resulted in high detection accuracy when segmented lung regions were used. The accuracy was raised from 97% to 99% when the segmentation of images was used. Researchers concluded that ChexNet outperforms other deep learning models whereas DenseNet201 outperforms segmented lungs (Rahman *et al.*, 2020).

Researchers developed a Deep Convolution Neural Network model using chest x-ray of TB patients to quantify the generalizability of the machine learning model. For these chest X-rays, the dataset is trained on one population and then tested with another population. The researchers conclude that diagnostic performance varies for a different population of datasets when trained using a supervised deep learning model (Sathitratanacheewin, Sunanta and Pongpirul, 2020a).

Researchers aimed to develop a computer-aided system that improves TB diagnostics using a convolution neural network. A deep learning model is also proposed for the localization in CXRs. The two publicly available CXR datasets, namely, the National Institutes of Health CXR dataset and the Shenzhen Hospital CXR dataset were used for this purpose. The whole process was performed following three important steps, namely the first CNN structures were modified, secondly, the artificial bee colony algorithm was used to fine-tune the model and the last linear average-based ensemble method was implemented. It is concluded that by superimposing the above three steps the overall performance of the Deep Convolution Neural Network increased (Guo, Passi and Jain, 2020).

Researchers proposed a system that can detect Tuberculosis consists of dual subsystems; one a data acquisition system where acquisition of field of views of sputum is automated, the second subsystem is a recognition system where transfer learning method (Inception V3 a DeepNet model) is implemented by customizing it and SVM classifies the data from transferred knowledge. An accuracy of 95.05% is attained and specificity and sensitivity are increased (Dinesh Jackson Samuel and Rajesh Kanna, 2019).

Researchers detected lung tuberculosis by proposing a Computer-Aided Diagnosis system that is based on image processing so that quick and accurate detection can be done. The proposed system consists of data cleaning, segmentation, feature extraction, and classification of data. It is concluded in this research that the system's specificity can be increased by maintaining accuracy and sensitivity. The system yields a sensitivity of 66.67%, specificity of 86%, and 76% of accuracy (Ilena, Kamarga and Setiawan, 2018).

Researchers propose a deep neural network that takes zoomed microscopy images of sputum and returns the location of TB bacilli as output. This proposed model gives high sensitivity with a recall of 83.78% and precision of 67.55% (Kant, 2018).

A system named TB-AI is built by researchers using convolution neural networks which are used to recognize acid-fast stained bacillus of TB specifically. 45 samples with 30 positive and 30 negative cases are included in the training dataset and 210 samples with 108 positive and 93 negative cases are collected as a test set. This system achieved sensitivity and specificity of 97.94% and 83.65% respectively (Xiong *et al.*, 2018).

Researchers proposed a method to classify and count the ZN-stained TB bacilli using the digital image processing technique where various steps are used. Like for bacilli segmentation, a hue color component-based approach is used. Similarly, shape characterization is used to declare whether the bacillus is valid or not. By presenting various images, it is shown that this method achieves an accuracy of 90% in counting the bacilli (Payasi and Patidar, 2018).

Researchers proposed an initial screening for TB infection using the non-linear SVM. Further, the gradient evolution algorithm is used to find the best parameter. The easily available data without containing any medical test is used for classification. In addition, the C5.0 decision tree is also employed for further analysis of TB infection. It is concluded that the proposed algorithm provides significant results with a small error rate (Zulvia, Ferani E.; Kuo, R.J.; Roflin, 2017).

Researchers presented a novel predictive algorithm to detect TB. This has been done by modifying the standard decision tree. This algorithm outperforms the other classifiers such as SVM, Random Forest; KNN resulted in 94% accuracy in detecting TB. Researchers conclude that relying on this kind of approach does not need any technical staff and provides significant results (Hussainy *et al.*, 2017).

Researchers identify Tuberculosis using Naïve Bayesian method. The data used in this research is taken from several health centers and various medical records of patients were used who were diagnosed with PTB and EPTB. Researchers concluded that Tuberculosis can be automatically detected well using the Naïve Bayesian method of machine learning and got the highest accuracy i.e. 85.95% using lab result feature and X-ray (Trihartati S. and Adi, 2016).

Researchers established the optimal decision tree model by using the BPS (Biomarker Patterns Software) which helps identify the EPTB and non-EPTB groups with the specificity of 97.7% and sensitivity of 84.1%. Also, this model achieved a classification accuracy of 91.6% which suggests that the decision tree model can play a potential role in diagnosing EPTB (Deng *et al.*, 2011).

Table 1: Comparison of Tuberculosis Detection Methods

Author and	Method	Dataset	Prediction	Approach
Year (Xing et al., 2020)	Support Vector Machine (SVM)	116 patients and 103 quantitative features were involved.	Provides AUC of 0.84, accuracy of 0.85 with sensitivity and specificity of 0.88 and 0.80 respectively.	Researchers explored machine learning approach to differentiate the non tuberculous mycobacteria (NTM) and PTB.
(Li, Huang and Jin, 2018)	Auto Encoder Convolution Neural Network	No dataset was mentioned.	81.26% accuracy was achieved with recall and fl score of 81.72% and 81.49% respectively.	The Auto Encoder Convolution Neural Network is used by researchers for the classification of PTB based on CT images.
(Li et al., 2021)	Convolution Neural Network	501 CT image dataset of active PTB and 501 for negative samples.	Model provides recall of 98.7% and precision of 93.7%.	Researchers establish a deep learning system that quantitative computed tomography (CT) reports for the diagnosis of PTB.
(Lakhani and Sundaram, 2017)	Two Deep Convolution Neural Networks; AlexNet and GoogleNet were used.	1007 posteroanterior chest radiographs	Model provides AUC of 0.99 with sensitivity of 97.3% and specificity 100%.	Researchers evaluate the efficacy of DCNNs for detection of TB on chest radiographs.
(Wu, Wang and Wu, 2017)	Random Forest was used.	485 PTB patients and 1990 Sarcoidosis patients data was used.	Model performed classification with AUC of 81%.	Researchers proposed the model that can differentiate between PTB and sarcoidosis.
(Ho et al., 2019)	Deep Convolution Neural Network; ResNet152, Inception-ResNet and DenseNet121 was used.	Public dataset ChestXray14 was used as training and two datasets Montgomery and Shenzhen were used as testing datasets.	Highest AUC achieved for DenseNet121 i.e. 0.95.	Researchers examine the efficiency of DCNNs for the detection of TB on chest radiographs.
(Trihartati S. and Adi, 2016)	Naïve Bayes classiier was used.	237 data samples were used.	Accuracy achieved was 85.95%.	The type of TB is identified by researches

			using Naïve Bayesian Classifier
(Tan et al., 2012)	No dataset was mentioned.	94.9% accuracy was achieved.	Computer-Assisted diagnosis of tuberculosis was done using first order statistical approach to chest radiographs.

3 Proposed System

3.1 Dataset

The dataset used in this research came from various health centers of the Jalandhar Region. It comprises basic parameters and site-related parameters. Basic parameters contain eight clinical parameters and six hematological parameters. Site-related parameters contain six spinal parameters, six pleural parameters, and three lymph node parameters. All the parameters and its values that have been selected for PTB and EPTB are taken from the literature survey and finalized after consulting with experts (Chen *et al.*, 2016), (Jung *et al.*, 2015), (Kurup *et al.*, 2016).

Researchers found that TB patients experienced cough, fever, loss of appetite, weight loss, night sweat (Saloman, 2016), (Van Crevel *et al.*, 2002); therefore these parameters are included in clinical parameters. Secondly, Hematological parameters are used as researchers find that hematological abnormalities are common in patients with TB (Abay *et al.*, 2018). These parameters mainly include HB, RBC, WBC, PLT, and Blood Sugar. Skin Test (Mantoux Tuberculin Skin Test (TST)) is also added as one parameter to confirm TB. This test was done by injecting a small amount of fluid into the skin and the result depends on the bump raised in the injected area. The patient with a bump size greater than 5mm has a high risk of TB, bump size with 10mm is at moderate risk of TB, and bump size greater than 15mm has no risk factors (Healthwise Staff, 2020), (Nayak and Acharjya, 2012). Table 2 shows the above parameters which are common in any TB patient. So, to make it confirm whether the patient has PTB, the patient's X-Ray is also considered as one parameter in detecting PTB.

Table 2: PTB Features

Name	Value
Cough (in weeks)	Greater than 3 weeks
Chest pain	Mild, moderate, severe
Hemoptysis (coughing up blood)	Yes, No
Fever (Fahrenheit)	Mild, moderate, severe
Fatigue	Yes, No
Weight loss	Yes, No
Loss of Appetite	Yes, No
Night Sweat	Mild, moderate, severe
HB (g /dL)	men: 13-17
	women: 12-15
RBC (cells/μL)	men: 4.7 to 6.1 million
	women: 4.2 to 5.4 million
WBC (cells/μL)	4500 to 11000

PLT (/μL)	150000-450000
Blood Sugar	<140 mg/dL
Skin test (bump size in mm)	bump < 5mm
X-Ray	Positive, Negative

EPTB can be seen in many different sites e.g. Lymph Node TB, Pleural TB, Bones and Joint TB, Central Nervous But in this research, 3 sites are chosen which are most common i.e. Lymph Node, Pleural, Spinal. About half of EPTB cases are comprised of TB in Lymph Nodes followed by Pleural TB which comprises about 40% (Gandhare and Mahashur, 2017a), (K.B. Gupta, Manju Jose, Uma Chaudhary, 2014). The involvement of Bone and Joint TB is about 10%, among those, spine is the most affected site as it is found almost in 50% of cases of Bone and Joint TB patients (Patel *et al.*, 2016), (Pigrau-Serrallach and Rodríguez-Pardo, 2013). Table 3 shows parameters for EPTB.

Table 3: EPTB Features

Name	Value
Swelling in the neck region	Yes, no
CRP level (mg/l)	<10mg/L
Cough (productive or not)	Yes, no
Pleuristic chest pain	Yes, no
Lymphocyte (%)	18 to 45% of WBC
ADA (U/ml)	< 40 U/ml
Protein level (g/L)	< 30 g/L
Severe back pain	Yes, no
Stiffness (inability to move easily)	Yes, no
Abscesses (swollen area with an accumulation of pus)	Yes, no
ESR (mm/hr)	men: 0 to 22
	women: 0 to 29
X-Ray Report	Positive, negative

Lymph Node TB is one of the most common types of EPTB. In studies conducted by experts or researchers, it is found that about half of the EPTB dataset contains patients with Lymph Node TB (Gandhare and Mahashur, 2017b). Swelling is the most common symptom in patients while cough and chest pain were less common. But other clinical symptoms like fever, night sweat, fatigue, weight loss are present (Maji *et al.*, 2015), (Amine Benjelloun, Youssef Darouassi, Yasser Zakaria, Rachid Bouchentouf, 2015), (Alihalassa, 2018), (Seok *et al.*, 2019), (Mathiasen *et al.*, 2020). C- Reactive Protein (CRP) is also considered one parameter because high CRP levels have high specificity for TB (Garcia-Pachon *et al.*, 2005).

Pleural TB is the second most common type of EPTB which is found in about 25% of EPTB patients (Jeon, 2014), (Light, 2015). Adenosine Deaminase (ADA), Lymphocytes, Protein levels are mainly used to detect its presence (Wright, 2008), (Valdés *et al.*, 2010). The easiest way to establish Pleural TB in patients is to demonstrate an ADA level above 40 U/L (Zhai, Lu and Shi, 2016).

Spinal TB is another form of EPTB that accounts for about 10 to 35% of cases (Stout, 2021). Chronic back pain is the only symptom was observed in 61% of cases of spinal tuberculosis. Other than back pain, common parameters include local pain, stiffness, and abscesses. Other symptoms include fever, night sweat, weight loss (Garg *et al.*, 2017), (Mousa, 2007). Researchers have suggested that TB is strongly associated with high ESR (Erythrocyte Sedimentation Rate) values that can be greater than 100 mm/hr, but the average ESR range for men is 0 to 22 mm/hr and for women 0 to 29 mm/hr (Clinic, 2019), (Mandal, 2016). This parameter is added in the Spinal TB parameter because ESR elevation is common in Spinal TB patients (Dharmalingam, 2004), (Mulleman *et al.*, 2006). This study showed that serum hemoglobin level, RBC count, and platelet count were reduced in tuberculosis patients and ESR, CRP, and WBC count were increased compared to healthy controls (Rohini *et al.*, 2016).

All the data used in this research is expert-generated descriptor features. The basic features such as cough, fever, weight loss, loss of appetite, among others, will remain common for both PTB and EPTB datasets. The PTB dataset contains these common features plus X-Rays of lungs, whereas EPTB contains the extra features depending upon the site of TB such for spinal TB back pain is the most common feature; similarly for pleural TB, the pleuritic chest pain is the common feature; same as for lymph node, the swelling in neck is common. Taking X-Rays of that particular site will confirm whether the patient has EPTB or not. The X-Ray images are used by extracting features out of them and then these features are combined with the expert-generated descriptor features.

The total number of patients that are included in this research is 249. Out of which, 200 samples are used to train the model and 49 samples are used to test the model. The training dataset consisted of 90 patients who had PTB, 67 patients who had EPTB, and the rest 43 patients had no TB.

3.2 Methodology

This research proposed an integrated TB detection system that can detect both types of TB i.e. PTB and EPTB using machine learning algorithms. For its implementation, the steps followed in this research are depicted in the flow chart given in Figure 1. First and foremost, the number of meetings has been organized with doctors/ clinicians, and a review of various pieces of literature was carried out to finalize the input variables (Chen *et al.*, 2016), (Jung *et al.*, 2015), (Kurup *et al.*, 2016). After the recommendations from doctors/ clinicians, 8 clinical parameters, 6 hematological parameters, and 12 EPTB parameters are finalized. Once the parameters have been finalized, the collection of data is started. It is collected by visiting various hospitals of the Jalandhar Region. This dataset includes the clinical data as well as the image data. This dataset that has been collected needs to be gone through preprocessing stage.

In the preprocessing stage, data cleaning and data transformation operations are mainly performed on the clinical data. In data cleaning, the main task is to remove the rows having missing values. In data transformation, some columns need to be transformed like in gender column converting all M and F to male and female respectively, in fever column converting numerical to categorical variables, etc. Similarly, the transformation of other columns is also done. For the preprocessing of X-Ray images, first, the normalization was applied so that the range of pixels of each image gets normal or familiar with each other. After normalization, PCA (Principal Component Analysis) is applied to reduce a large number of image vectors.

After passing through the preprocessing stage, the data is divided into training and test data, including images. To convert these images into feature vectors, Image Embedder is used. Here, the feature vectors are calculated using deep learning models. In this research, SqueezeNet (Tsang, 2019), a deep learning model is used as it is a fast and small model for image recognition. This will add additional columns to the dataset which are large in number and therefore these vectors are reduced using PCA in this step. Now, it becomes easy for a machine learning model to train this dataset. Now we are dealing with two datasets: general dataset second is an image dataset, so we have to merge both datasets. It is now the turn to train the machine learning models.

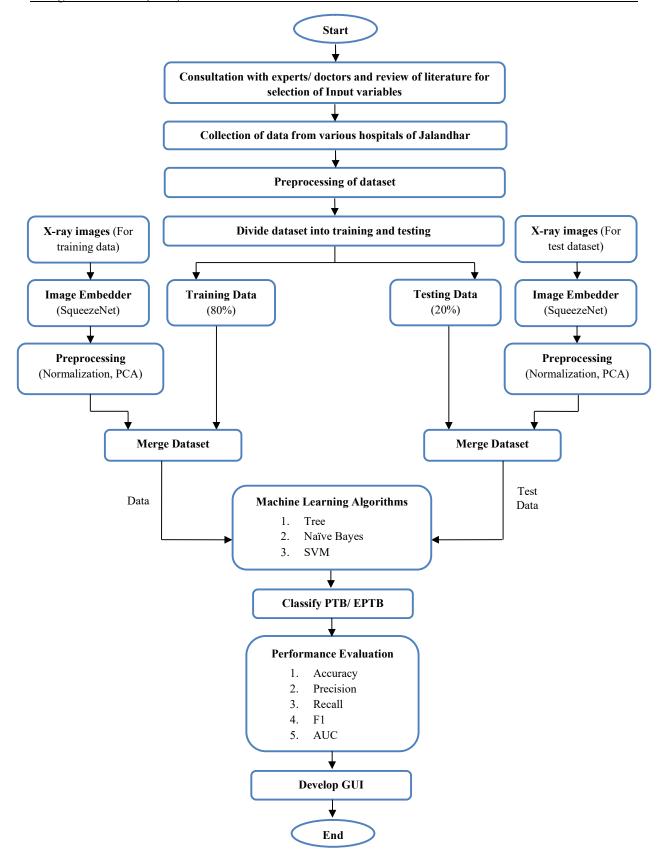


Figure 1: Flow chart for classification of PTB and EPTB

Researchers surveyed different machine learning algorithms for diagnosing different diseases and concluded that SVM had improved accuracy followed by Naïve Bayes. However, Decision Tree and Random Forest have also shown superior accuracy (Fatima and Pasha, 2017), (Uddin *et al.*, 2019). Therefore, this research chooses three basic algorithms: Decision Tree, Naïve Bayes, and Support Vector Machine.

The decision tree is a classification algorithm that represents the data in a tree-like structure. It is used to train the model so that it can be used to predict the target variable by learning decision rules. Some attributes need to be selected for producing decision tree. These attributes are Information Gain, Entropy etc (Akshay, 2021). Entropy is the measure of impurity and is given as:

$$Entropy(S) = \sum_{i=1}^{n} p_i * \log_2(p_i)$$

Information Gain is the measure of decrease in entropy after splitting and is given as

$$Information\ Gain(S, a) = Entropy(S) - Entropy(S|a)$$

where S is a sample and a is an attribute. Entropy(S) is the entropy of a sample before any change and Entropy(S|a) is conditional entropy for the given sample.

The Naïve Bayes is a classification algorithm that is based on the Bayes Theorem. This algorithm can predict the future opportunities using the probability and statistical approach. The equation of Bayes Theorem is given as under: (Trihartati S. and Adi, 2016)

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

where P(X|Y) is Liklihood which means how probable is the evidence given that the hypothesis is true.

P(Y) is Prior which means how probable was our hypothesis before observing the evidence.

P(X) is Marginal which means how probable is the new evidence under all possible hypotheses.

P(Y|X) is Posterior which means how probable is our hypothesis given the observed evidence.

The Support Vector Machine is another simple machine learning algorithm used for classification. Ot produces significant accuracy. Its main goal is to find the hyperplane with maximum margin that can classify the data points (Gandhi, 2018). The RBF (Radial Basis Function) is used as kernel function that is used to find the distance between data points. It is represented as (Sushant Sreenivasa, 2020):

$$K(X_1, X_2) = \exp(-\frac{\|X_1, X_2\|^2}{2\sigma^2})$$

where σ is the variance and $||X_1, X_2||$ is the Euclidean Distance of two points X_1 and X_2 .

The training dataset of 200 samples is given to all three algorithms and the other 49 samples are used for testing the models. After training and testing are done, the model will become able to classify Pulmonary Tuberculosis from Extra Pulmonary Tuberculosis.

After passing through the modeling stage, various performance evaluation parameters are calculated, including Confusion Matrix, the Area under ROC curve, the Classification Accuracy, the Precision, the Recall, and the F1 Score.

In the last stage, the Graphical User Interface will be made to make it easy for doctors/clinicians to go through the system. After that validation of this system will be done by experts/doctors.

4 Results

In this study, classification of both types of tuberculosis is done with the help of three basic machine learning algorithms i.e. Decision Tree, Naïve Bayes, and Support Vector Machine. The performance is evaluated based on five parameters i.e. Area under ROC Curve or AUC value, Classification Accuracy, Precision, Recall, and F1 value.

During experimentation, a certain amount of data is set aside to test the final performance of a trained model, while the rest of the database is divided into a training database. The training and testing dataset is set into 80:20 ratios. With the training dataset, the model got trained and for the rest of the data, testing is performed. Then, the models are evaluated.

The first performance evaluation parameter is AUC which is Area under the ROC curve that gives a graph between True Positive Rate and False Positive Rate which can be calculated as under (Sarang Narkhede, 2019). It tells how capable this model is to correctly predict the instances. Higher the AUC value better will be the model to distinguish between PTB, EPTB, and No TB patients. Below are the equations using which the AUC is calculated:

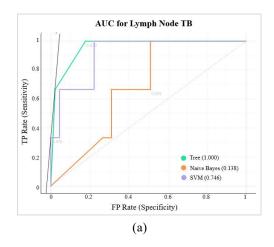
$$TPR = 100 * (\frac{TP}{TP + FN})$$

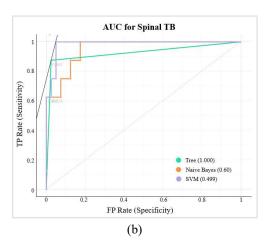
$$FPR = 100 * (\frac{FP}{FP + TN})$$

where TP is a True Positive number which predicts the patient as having TB when he or she is actually having TB and TN is a True Negative number which predicts the patient as not having TB when he's is actually not having any TB. Thant means TP and TN tell the correct number of classifications. FP is a False positive number that predicts the patient as having TB when he or she is not actually having TB, and FN is a False Negative number that predicts the patient as not having TB when he or she is actually having. That means FP and FN are errors and not classifying correctly.

Table 4 reveals that after comparing the three algorithms, the AUC value for the Decision Tree is 0.95 which is the highest when compared to the other two algorithms with Naïve Bayes having 0.84 and SVM having 0.91.

It is evident from the Figure 2(a), 2(c), 2(d), 2(e) that the AUC for the Decision Tree ROC curve is higher than that for the SVM and Naïve Bayes ROC curves while Figure 2(b) shows something different as here ROC curve of SVM is higher than that of Decision Tree and Naïve Bayes. Therefore, we can say that the Decision Tree did a better job for classifying the Lymph Node TB, Pleural TB, Lungs TB, and No TB cases while in the case of Spinal TB SVM performs better.





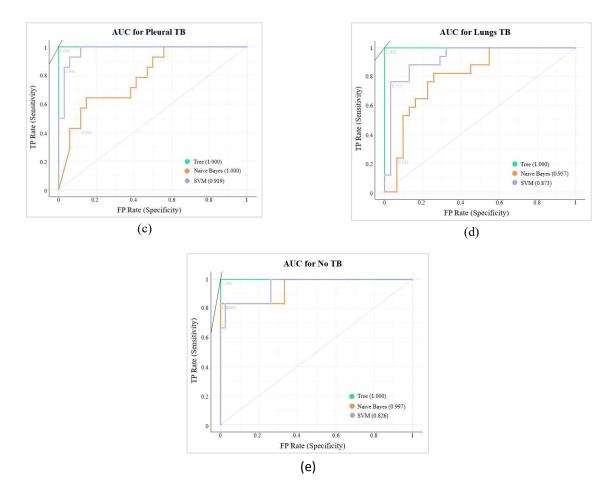


Figure 2: (a) AUC for Lymph Node TB, (b) AUC for Spinal TB, (c) AUC for Pleural TB, (d) AUC for Lungs TB, (e) AUC for No TB

Other than AUC value, Accuracy, Precision, Recall, F1 Scores are also calculated using the formulas below:

Accuracy: The classification accuracy is used to measure the effectiveness of the proposed method and to compare the three basic algorithms used in this research.

$$Accuracy = \frac{Number\ of\ correct\ TB\ predictions}{Total\ number\ of\ TB\ predictions}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is used to measure the proportion of true positives among instances classified as positive. It tells the measure of patients that we correctly identify having TB out of all the patients actually having it.

$$Precision = \frac{correctly\ predicted\ positive\ TB\ cases}{total\ number\ of\ positive\ cases\ that\ were\ predicted}$$

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall is used to measure the proportion of true positives among all positive instances in the data. For all those who actually have TB, recall tells us how many we correctly identified as having TB.

$$Recall = \frac{correctly\ predicted\ positive\ TB\ cases}{total\ number\ of\ actual\ positive\ cases}$$

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: The F1-score is used as a single measure that combines both precision and recall.

$$F1 \, Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$F1 \, Score = \frac{2TP}{2TP + FP + FN}$$

It is evident from Table 4 that the Tree is giving the best Accuracy, Precision, Recall, F1 Score with 98%, 95%, 95%, 95% respectively while Naïve Bayes is giving the worst.

Table 4. I chomiance marces						
Model	AUC	CA	F1	Precision	Recall	
Tree	0.98	0.95	0.96	0.97	0.95	
Naïve Bayes	0.83	0.89	0.87	0.91	0.89	
SVM	0.95	0.91	0.90	0.92	0.91	

Table 4: Performance Indices

Figure 3 shows that amongst the classification models, it can be noticed that the Decision Tree outperforms the other classifiers in terms of all performance evaluation parameters while Naïve Bayes performs lowest for the input dataset.

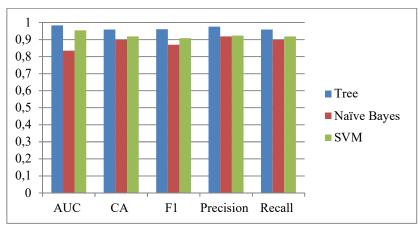


Figure 3: Performance of all algorithms

Confusion Matrix: To see the performance of classification with more reliability, a confusion matrix is used as it describes all measures that how many are correctly predicted when it actually true or how many cases are predicted as true when it is not actually, etc. It compares the actual TB patients with those predicted by machine learning

models. Confusion Matrix is evaluated in two ways: (a) based on the EPTB site shown in Table 5, 6, 7 (b) based on the type of TB shown in Table 8, 9, 10. Confusion Matrix based on the classification of EPTB Site:

Confusion Matrix based on the classification of EPTB Site:

Table 5: Confusion Matrix for Decision Tree

		Predicted					
		Lungs TB	Lymph Node TB	No TB	Pleural TB	Spinal TB	
	Lungs TB	18	0	0	0	0	
=	Lymph Node TB	0	2	0	0	1	
Actual	No TB	0	0	6	0	0	
▼	Pleural TB	0	0	0	14	0	
	Spinal TB	0	1	0	0	7	

Table 6: Confusion Matrix for Naïve Bayes

		Predicted					
		Lungs TB	Lymph Node TB	No TB	Pleural TB	Spinal TB	
	Lungs TB	2	11	0	5	0	
귵	Lymph Node TB	0	2	0	0	1	
Actual	No TB	1	0	5	0	0	
V	Pleural TB	1	7	0	6	0	
	Spinal TB	0	3	0	0	5	

Table 7: Confusion Matrix for SVM

		Predicted					
		Lungs TB	Lymph Node TB	No TB	Pleural TB	Spinal TB	
	Lungs TB	16	0	1	1	0	
₌	Lymph Node TB	1	1	0	0	1	
Actual	No TB	1	0	5	0	0	
▼	Pleural TB	4	0	0	10	0	
	Spinal TB	0	0	0	0	8	

Confusion Matrix based on the classification of the type of TB:

Predicted EPTB No TB PTB **EPTB** 22 0 1 Actual No TB 3 6 1 PTB 0 0 16

Table 8: Confusion Matrix for Decision Tree

Table 9: Confusion Matrix for Naïve Bayes

		Predicted					
		ЕРТВ	No TB	PTB			
	ЕРТВ	20	2	3			
Actual	No TB	0	5	1			
	PTB	4	4	10			

Table 10: Confusion Matrix for SVM

		Predicted					
		ЕРТВ	No TB	PTB			
_	ЕРТВ	23	0	2			
Actual	No TB	0	5	1			
Į.	PTB	2	1	15			

5 Discussion

There are various ways to diagnose TB, such as augmentation, sputum release, and other tests; however, these methods take longer to process results, require special types of equipment, and are expensive. Researchers also report that conventional methods were reported to be less than 50% accurate in some cases when TB was detected. Therefore they are always looking for new ways that can provide an early diagnosis with good accuracy (Graham *et al.*, 2012). Furthermore, researchers have become interested in using artificial intelligence and machine learning to diagnose TB (Shamshirband *et al.*, 2014). Classification is also becoming the most important decision-making tool in medicine (Kahramanli and Allahverdi, 2008) (Tunc, 2012). Therefore, machine learning technology is used in this research to classify tuberculosis as it is good in finding patterns from historical data and can detect the disease more accurately.

This study investigated both the diagnosis of Pulmonary tuberculosis and Extra Pulmonary Tuberculosis. To detect both types of TB, in addition to the accuracy and AUC score, the most commonly used methods include precision, recall, and F1 score that have been used to provide a complete analysis of the models. Among the three basic machine learning models (Decision Tree, Naïve Bayes, and SVM) used in this study, the decision tree obtained the best accuracy of 95%, AUC value of 0.98, Precision of 0.95, Recall of 0.95, and F1 score of 0.95. The results not only reflect the AUC score of the three models but also offer comprehensive measures, namely Accuracy, Precision, Recall, and F1 score for all models.

Figure 3 shows a comparison of machine learning models on the collected dataset. Results show that the Decision Tree outperforms the other classifiers in terms of performance evaluation parameters while Naïve Bayes performs the lowest for the input dataset.

To the best of our knowledge, research on an integrated PTB and EPTB acquisition model is lacking. In this area, this work provides a baseline in diagnosing both types of TB using machine learning methods. Unlike most of the research work that focused solely on PTB, this work provides comprehensive prediction approaches to PTB and EPTB detection.

Although it might be reasonable to say that the decision tree provides the best accuracy compared to other algorithms, there are always many other factors that can influence the algorithm's performance. So, the authors can do further research on the simulation of TB cases is recommended.

6 Conclusion

It is a challenging task to identify TB due to the occurrence of complex patterns present in CXRs. The literature already proposed several approaches to identify PTB but low attention is given to the identification of EPTB using machine learning models. The proposed methodology can classify both types of TB i.e. PTB and EPTB. The results have demonstrated that the proposed methodology achieved significant results in terms of AUC, Accuracy, Precision, Recall, F1 Score. After comparing three basic algorithms, results show that the Decision Tree outperforms the other two algorithms i.e. Naïve Bayes and SVM. This proposed work provides a good foundation for future work. In the future, this work can be extended using large datasets. Furthermore, advanced machine learning algorithms can be used for more accurate results.

References

Abay, F. *et al.* (2018) 'Hematological Abnormalities of Pulmonary Tuberculosis Patients with and without HIV at the University of Gondar Hospital, Northwest Ethiopia: A Comparative Cross-Sectional Study', *Tuberculosis Research and Treatment*, 2018, pp. 1–6. doi: 10.1155/2018/5740951.

Akshay, S. (2021) *Decision Tree Algorithm for Classification: Machine Learning 101, Analytic Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/(Accessed: 30 September 2021).

Alihalassa, S. (2018) 'Lymph node tuberculosis: Criteria for cure and optimal duration of treatment', *European Respiratory Journal*, 52. doi: https://doi.org/10.1183/13993003.congress-2018.PA2691.

Amine Benjelloun, Youssef Darouassi, Yasser Zakaria, Rachid Bouchentouf, N. E. (2015) 'Lymph nodes tuberculosis: a retrospective study on clinical and therapeutic features', *Pan African Medical Journal*, 8688, pp. 1–3. doi: 10.11604/pamj.2015.20.65.5782.

Chen, C. H. et al. (2016) 'Early diagnosis of spinal tuberculosis', Journal of the Formosan Medical Association. Elsevier Ltd, 115(10), pp. 825–836. doi: 10.1016/j.jfma.2016.07.001.

Clinic, M. (2019) Sed rate (erythrocyte sedimentation rate), Mayo Clinic. Available at: https://www.mayoclinic.org/tests-procedures/sed-rate/about/pac-20384797 (Accessed: 29 March 2021).

Van Crevel, R. *et al.* (2002) 'Decreased plasma leptin concentrations in tuberculosis patients are associated with wasting and inflammation', *Journal of Clinical Endocrinology and Metabolism*, 87(2), pp. 758–763. doi: 10.1210/jcem.87.2.8228.

Deng, C. et al. (2011) 'Establishing a serologic decision tree model of extrapulmonary tuberculosis by MALDI-TOF MS analysis &, &&;', Diagnostic Microbiology and Infectious Disease. Elsevier Inc., 71(2), pp. 144–150. doi: 10.1016/j.diagmicrobio.2011.06.021.

Dharmalingam, M. (2004) 'Tuberculosis of the spine - The Sabah experience. Epidemiology, treatment and results', *Tuberculosis*, 84(1–2), pp. 24–28. doi: 10.1016/j.tube.2003.08.008.

Dinesh Jackson Samuel, R. and Rajesh Kanna, B. (2019) 'Tuberculosis (TB) detection system using deep neural networks', *Neural Computing and Applications*. Springer London, 31(5), pp. 1533–1545. doi: 10.1007/s00521-018-3564-4.

Fatima, M. and Pasha, M. (2017) 'Survey of Machine Learning Algorithms for Disease Diagnostic', *Journal of Intelligent Learning Systems and Applications*. Scientific Research Publishing, Inc, 09(01), pp. 1–16. doi: 10.4236/jilsa.2017.91001.

Gandhare, A. and Mahashur, A. (2017a) 'Tuberculosis of the lymph nodes: Many facets, many hues', *Astrocyte*. Medknow, 4(2), p. 80. doi: 10.4103/astrocyte.astrocyte 65 17.

Gandhare, A. and Mahashur, A. (2017b) 'Tuberculosis of the lymph nodes: Many facets, many hues', *Astrocyte*. Medknow, 4(2), p. 80. doi: 10.4103/astrocyte.astrocyte_65_17.

Gandhi, R. (2018) Support Vector Machine — Introduction to Machine Learning Algorithms, Towards Data Science. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed: 30 September 2021).

Garcia-Pachon, E. *et al.* (2005) 'C-reactive protein in lymphocytic pleural effusions: A diagnostic aid in tuberculous pleuritis', *Respiration*, 72(5), pp. 486–489. doi: 10.1159/000087672.

Garg, R. K. et al. (2017) 'Spinal tuberculosis: A review', *The Journal of Spinal Cord Medicine*, 0268(November). doi: 10.1179/2045772311Y.0000000023.

Graham, S. M. *et al.* (2012) 'Evaluation of Tuberculosis Diagnostics in Children: 1 . Proposed Clinical Case Definitions for Classification of Intrathoracic Tuberculosis Disease . Consensus From an Expert Panel', *The Journal of Infectious Diseases*, 205(Suppl 2), pp. 199–208. doi: 10.1093/infdis/jis008.

Guo, R., Passi, K. and Jain, C. K. (2020) 'Tuberculosis Diagnostics and Localization in Chest X-Rays via Deep Learning Models', *Frontiers in Artificial Intelligence*, 3(October), pp. 1–17. doi: 10.3389/frai.2020.583427.

Healthwise Staff (2020) *Tuberculin Skin Test*, *Michigan Medicine*. Available at: https://www.uofmhealth.org/health-library/hw203560#hw203602 (Accessed: 29 March 2021).

Ho, T. K. K. et al. (2019) Utilizing Pretrained Deep Learning Models for Automated Pulmonary Tuberculosis Detection Using Chest Radiography, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing. doi: 10.1007/978-3-030-14802-7_34.

Hussainy, S. et al. (2017) 'Decision-tree inspired classification algorithm to detect Tuberculosis (TB)', in *Pacific Asia Conference on Information Systems (PACIS)*. Available at: https://aisel.aisnet.org/pacis2017/182 (Accessed: 24 September 2021).

Ilena, G., Kamarga, S. A. and Setiawan, A. W. (2018) 'Early Detection of Tuberculosis using Chest X-Ray (CXR) with Computer-Aided Diagnosis', *Proceedings of 2018 2nd International Conference on Biomedical Engineering: Smart Technology for Better Society, IBIOMED 2018.* IEEE, pp. 76–79. doi: 10.1109/IBIOMED.2018.8534784.

Jeon, D. (2014) 'Tuberculous pleurisy: An update', *Tuberculosis and Respiratory Diseases*, 76(4), pp. 153–159. doi: 10.4046/trd.2014.76.4.153.

Jung, Y. J. *et al.* (2015) 'The significance of sensitive interferon gamma release assays for diagnosis of latent tuberculosis infection in patients receiving tumor necrosis factor-aα antagonist therapy', *PLoS ONE*, 10(10), pp. 1–12. doi: 10.1371/journal.pone.0141033.

K.B. Gupta, Manju Jose, Uma Chaudhary, S. M. (2014) 'Pulmonary involvement in extrapulmonary tuberculosis patients', *European Respiratory Journal*, 38.

Kahramanli, H. and Allahverdi, N. (2008) 'Design of a hybrid system for the diabetes and heart diseases', *Expert Systems with Applications*, 35, pp. 82–89. doi: 10.1016/j.eswa.2007.06.004.

Kant, S. (2018) 'Towards Automated Tuberculosis detection', 2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, pp. 1250–1253.

Kurup, R. et al. (2016) 'Hematological and Biochemistry Profile and Risk Factors Associated with Pulmonary Tuberculosis Patients in Guyana', *Tuberculosis Research and Treatment*, 2016, pp. 1–6. doi: 10.1155/2016/6983747.

Lakhani, P. and Sundaram, B. (2017) 'Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks', *Radiology*, 284(2), pp. 574–582. Available at: http://pubs.rsna.org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1148/radiol.2017162326.

Li, L., Huang, H. and Jin, X. (2018) 'AE-CNN Classification of Pulmonary Tuberculosis Based on CT Images', in *Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018*. IEEE, pp. 39–42. doi: 10.1109/ITME.2018.00020.

Li, X. et al. (2021) 'A deep learning system that generates quantitative CT reports for diagnosing pulmonary Tuberculosis', *Applied Intelligence*. Applied Intelligence, 51(6), pp. 4082–4093. doi: 10.1007/s10489-020-02051-1.

Light, R. W. (2015) 'Tuberculous pleural effusion', Turk Toraks Dergisi, 16(1), pp. 1–9. doi: 10.5152/ttd.2014.001.

Maji, A. *et al.* (2015) 'Expression profiling of lymph nodes in tuberculosis patients reveal inflammatory milieu at site of infection', *Scientific Reports*. Nature Publishing Group, 5(September), pp. 1–10. doi: 10.1038/srep15214.

Mandal, S. K. (2016) 'Erythrocyte Sedimentation Rate Values in Cases of Active Tuberculosis without HIV Co-Infection', *Journal of Medical Science And clinical Research*, 04(10), pp. 13156–13159. doi: 10.18535/jmscr/v4i10.58.

Mathiasen, V. D. *et al.* (2020) 'Clinical features of tuberculous lymphadenitis in a low-incidence country', *International Journal of Infectious Diseases*. International Society for Infectious Diseases, 98, pp. 366–371. doi: 10.1016/j.ijid.2020.07.011.

Mousa, H. A. L. (2007) 'Bones and joints tuberculosis', *Bahrain Medical Bulletin*, 29(1), pp. 17–21. doi: 10.5035/nishiseisai.25.333.

Mulleman, D. et al. (2006) 'Characteristics of patients with spinal tuberculosis in a French teaching hospital', *Joint Bone Spine*, 73(4), pp. 424–427. doi: 10.1016/j.jbspin.2005.10.013.

Nayak, S. and Acharjya, B. (2012) 'Mantoux test and its interpretation', *Indian Dermatology Online Journal*, 3(1), p. 2. doi: 10.4103/2229-5178.93479.

Patel, R. et al. (2016) 'Spinal Tuberculosis and Cold Abscess without Known Primary Disease: Case Report and Review of the Literature', Case Reports in Infectious Diseases. Hindawi Limited, 2016, pp. 1–4. doi: 10.1155/2016/1780153.

Payasi, Y. and Patidar, S. (2018) 'Diagnosis and counting of tuberculosis bacilli using digital image processing', in *IEEE International Conference on Information, Communication, Instrumentation and Control, ICICIC 2017*, pp. 1–

5. doi: 10.1109/ICOMICON.2017.8279128.

Pigrau-Serrallach, C. and Rodríguez-Pardo, D. (2013) 'Bone and joint tuberculosis', *European Spine Journal*, 22(SUPPL.4). doi: 10.1007/s00586-012-2331-y.

Prakasha, S. R. et al. (2013) 'Mapping the pattern and trends of extrapulmonary tuberculosis', *Journal of Global Infectious Diseases*. Wolters Kluwer -- Medknow Publications, 5(2), pp. 54–59. doi: 10.4103/0974-777X.112277.

Purohit, M. and Mustafa, T. (2015) 'Laboratory diagnosis of extra-pulmonary tuberculosis (EPTB) in resource-constrained setting: State of the art, challenges and the need', *Journal of Clinical and Diagnostic Research*, 9(4), pp. EE01–EE06. doi: 10.7860/JCDR/2015/12422.5792.

Rahman, T. et al. (2020) 'Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization', *IEEE Access*, 8, pp. 191586–191601. doi: 10.1109/ACCESS.2020.3031384.

Rohini, K. et al. (2016) 'Assessment of Hematological Parameters in Pulmonary Tuberculosis Patients', *Indian Journal of Clinical Biochemistry*. Springer India, 31(3), pp. 332–335. doi: 10.1007/s12291-015-0535-8.

Saloman, L. (2016) Examining Disease Characteristics and Patterns of Weight Gain in Tuberculosis Treatment, Contagion Live. Available at: https://www.tbonline.info/posts/2016/11/12/examining-disease-characteristics-and-patterns-wei/ (Accessed: 29 March 2021).

Sarang Narkhede (2019) 'Understanding AUC - ROC Curve', *Towards Data Science*, pp. 6–11. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: 1 July 2021).

Sathitratanacheewin, S., Sunanta, P. and Pongpirul, K. (2020a) 'Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability', *Heliyon*. Elsevier Ltd, 6(8), p. e04614. doi: 10.1016/j.heliyon.2020.e04614.

Sathitratanacheewin, S., Sunanta, P. and Pongpirul, K. (2020b) 'Heliyon Deep learning for automated classi fi cation of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability', *Heliyon*. Elsevier Ltd, 6(January), p. e04614. doi: 10.1016/j.heliyon.2020.e04614.

Seok, H. *et al.* (2019) 'Correction to: Characteristics of residual lymph nodes after six months of antituberculous therapy in HIV-negative individuals with cervical tuberculous lymphadenitis (BMC Infectious Diseases (2019) 19 (867) DOI: 10.1186/s12879-019-4507-0)', *BMC Infectious Diseases*. BMC Infectious Diseases, 19(1), pp. 1–7. doi: 10.1186/s12879-019-4640-9.

Shamshirband, S. et al. (2014) 'Tuberculosis Disease Diagnosis Using Artificial Immune Recognition System', *International Journal of Medical Sciences*, 11(5), pp. 508–514. doi: 10.7150/ijms.8249.

Shrivastava, D. A. K. et al. (2015) 'Clinico-Epidemiological Profile of Extra-pulmonary Tuberculosis in Central India', *International Journal of Medical Research and Review*, 3(2), pp. 223–230. doi: 10.17511/ijmrr.2015.i2.046.

Stout, J. (2021) *Bone and joint tuberculosis*, *UpToDate*. Available at: https://www.uptodate.com/contents/bone-and-joint-tuberculosis (Accessed: 29 March 2021).

Sushant Sreenivasa (2020) Radial Basis Function (RBF) Kernel: The Go-To Kernel, Towards Data Science. Available at: https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a (Accessed: 30 September 2021).

Tan, J. H. et al. (2012) 'Computer-Assisted Diagnosis of Tuberculosis: A First Order Statistical Approach to Chest Radiograph', *Journal of Medical Systems*, 36(5), pp. 2751–2759. doi: 10.1007/s10916-011-9751-9.

TOI (2020) 'Number of extrapulmonary TB cases nearly doubles in 3 yrs | Pune News - Times of India', *Times Of India*. Available at: https://timesofindia.indiatimes.com/city/pune/number-of-extrapulmonary-tb-cases-nearly-doubles-in-3-yrs/articleshow/73091447.cms (Accessed: 15 October 2020).

Trihartati S., A. and Adi, C. K. (2016) 'An Identification of Tuberculosis (Tb) Disease in Humans using Naive Bayesian Method', *Scientific Journal of Informatics*, 3(2), pp. 99–108. doi: 10.15294/sji.v3i2.7918.

Tsang, S.-H. (2019) *Review: SqueezeNet (Image Classification)*, *Towards Data Science*. Available at: https://towardsdatascience.com/review-squeezenet-image-classification-e7414825581a (Accessed: 8 November 2021).

Tunc, T. (2012) 'A New Hybrid Method Logistic Regression and Feedforward Neural Network for Lung Cancer Data', *Mathematical Problems in Engineering*, p. 10. doi: 10.1155/2012/241690.

Uddin, S. et al. (2019) 'Comparing different supervised machine learning algorithms for disease prediction', BMC Medical Informatics and Decision Making. BioMed Central Ltd, 19(1), pp. 1–16. doi: 10.1186/s12911-019-1004-8.

Ul Abideen, Z. et al. (2020) 'Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks', *IEEE Access*. IEEE, 8, pp. 22812–22825. doi: 10.1109/ACCESS.2020.2970023.

Valdés, L. *et al.* (2010) 'Diagnosing tuberculous pleural effusion using clinical data and pleural fluid analysis. A study of patients less than 40 years-old in an area with a high incidence of tuberculosis', *Respiratory Medicine*, 104(8), pp. 1211–1217. doi: 10.1016/j.rmed.2010.02.025.

Velingker, A., Lawande, D. and Dcosta, L. (2018) 'Section: Pulmonary Medicine Clinico-Epidemiological Profile of Extra Pulmonary Tuberculosis in Western India Section: Pulmonary Medicine', *International Journal of Contemporary Medical Research*, 5(2), pp. 2–5.

WHO (2020) 'Global Tuberculosis Report', World Health Organization.

Wright, J. L. (2008) 'Inflammatory and Fibrosing Pleural Processes', *Pulmonary Pathology*, pp. 733–744. doi: 10.1016/b978-0-443-06741-9.50041-0.

Wu, Y., Wang, H. and Wu, F. (2017) 'Automatic Classification of Pulmonary Tuberculosis and Sarcoidosis based on Random Forest', in 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics.

Xing, Z. et al. (2020) 'Machine Learning-Based Differentiation of Nontuberculous Mycobacteria Lung Disease and Pulmonary Tuberculosis Using CT Images', *BioMed Research International*, 2020. doi: 10.1155/2020/6287545.

Xiong, Y. et al. (2018) 'Automatic detection of mycobacterium tuberculosis using artificial intelligence', *Journal of Thoracic Disease*, 10(3), pp. 1936–1940. doi: 10.21037/jtd.2018.01.91.

Zhai, K., Lu, Y. and Shi, H. (2016) 'Tuberculous pleural effusion', *Journal of Thoracic Disease*, 8(4), pp. 486–494. doi: 10.21037/jtd.2016.05.87.

Zulvia, Ferani E.; Kuo, R.J.; Roflin, E. (2017) 'An Initial Screening Method for Tuberculosis Diseases Using a Multi-objective Gradient Evolution- based Support Vector Machine and C5 . 0 Decision Tree', in *IEEE 41st Annual Computer Software and Applications Conference*, pp. 204–209. doi: 10.1109/COMPSAC.2017.57.