Faculty of Natural and Applied Sciences Journal of Mathematical and Statistical Computing
Print ISSN: 3026-8605, e-ISSN 3043-632X
www.fnasjournals.com
Volume 2; Issue 2; March 2025; Page No. 8-14.

Prediction of Tuberculosis Using Machine Learning Techniques

***Olojido, J.B., & Oladimeji, B. J,**
Department of Computer Science, Rufus Giwa Polytechnic, Owo, Nigeria

***Corresponding author email:** olojidoj@gmail.com

Abstract
An early and precise diagnosis is necessary to improve treatment outcomes for tuberculosis (TB), which continues to be a major worldwide health concern. This study explores the application of machine learning algorithms to predict tuberculosis from chest X-ray images. By employing advanced techniques like Random Forests, Support Vector Machines (SVM), and Convolutional Neural Networks (CNNs), we created models that could identify TB-related anomalies in X-rays. Data pre-processing techniques like normalization and augmentation improved model performance, and a large dataset of labelled chest X-ray images was used. the proportion of samples that the model accurately separates from samples that are not TB-infected. This was achieved by leveraging the characteristics of True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN) from the confusion matrix. The models were assessed using various metrics, including accuracy, precision, recall, and F1 score, to ensure their robustness and dependability. Standard performance metrics such as accuracy, recall, precision, F1 score, and false alarm rate were utilized for evaluation, demonstrating the effectiveness of the proposed model, which achieved an accuracy of 93.0%, precision of 92.0%, recall of 94.0%, F1 score of 92.99%, and a false alarm rate of 0.078. Results indicate that machine learning can significantly aid in the early detection of TB, potentially leading to timely interventions. This research highlights the necessity of integrating machine learning tools into clinical practice to improve diagnostic accuracy and ultimately contribute to global TB control efforts.

Keywords: Tuberculosis Prediction, Machine Learning, Disease Diagnosis, Predictive Modeling, Healthcare Analytics

Introduction

Tuberculosis (TB) was once thought to be nearly eradicated, but it has resurfaced as a significant global issue. The disease is caused by a bacterium known as Mycobacterium tuberculosis. It can be transmitted between individuals, and those afflicted with tuberculosis may face fatal consequences if they do not receive appropriate treatment. This microorganism widely exists in humans, cattle, sheep, and birds. All of the organs in the body can be affected by tuberculosis. However, most of the tuberculosis cases occur in the lungs. Tuberculosis disease occurs under different manifestations in adults and children. When a person first comes into contact with the bacillus, typically during childhood, the lymphatic glands situated at the lung's entry point become the initial site of infection for this microorganism. Consequently, these glands swell (hilar lymphadenopathy), leading to what is referred to as primary tuberculosis. (Alemu, 2018)

The presence of microorganisms in phlegm must be demonstrated in order to make an accurate diagnosis. However, under a microscope, certain other microbes can also be identified as Mycobacterium tuberculosis. A unique culture medium is created where only Mycobacterium tuberculosis bacteria can proliferate in order to circumvent this issue. The patient's sputum sample is placed in this medium and incubated at body temperature for 45 days. After this duration, the culture medium is examined for any signs of bacterial growth. To treat tuberculosis, a combination of 4-5 different primary anti-tuberculosis antibiotics is administered over 6-12 months. In some instances, individuals may recover without any treatment if their immune system is sufficiently strong. However, even after complete recovery, lung damage caused by tuberculosis remains as calcified tissue. Unfortunately, untreated cases can lead to the patient's death. A period of 45 days is necessary to achieve an accurate diagnosis. This study aims to create a data mining solution that enhances the accuracy of tuberculosis diagnosis and assists in determining whether it is appropriate to initiate treatment for suspected patients without waiting for definitive test results. (Davidson, 2009).

Data mining refers to the process of discovering patterns in large data sets through methods that combine database systems, statistics, and machine learning techniques. It is a crucial procedure in which data patterns are extracted using clever techniques. It is a branch of computer science that is multidisciplinary. The primary goal of the data mining process is to extract information from a data set and transform it into a format that is understandable for future use. This process encompasses various stages, including database and data management, data pre-processing, model and inference considerations, interestingness metrics, complexity assessments, post-processing of identified structures, visualization, and online updates, in addition to the initial analysis phase. Data mining is referred to as the analytical component of the "knowledge discovery in databases" (KDD) process. The term can be misleading, as data mining specifically pertains to the analysis step within the KDD framework. It is a misnomer because the focus is on uncovering patterns and knowledge from extensive data sets rather than merely extracting the data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support systems, including artificial intelligence, machine learning, and business intelligence. (Tamer, 2011).

In machine learning, feature selection is a fundamental concept that significantly affects your model's performance. The features you choose for training your models play a crucial role in the results you can achieve. Therefore, feature selection and scaling should be prioritized as the initial and most critical steps in your model development process. Feature selection involves the automatic or manual identification of the features that have the greatest impact on your target variable or output of interest. Including irrelevant features in your dataset can reduce model accuracy and lead to learning based on non-essential attributes. It's important to note that feature selection and scaling differ from dimensionality reduction. While both approaches aim to decrease the number of features in a dataset, dimensionality reduction creates new combinations of features, whereas feature selection simply includes or excludes existing features without altering them. (Jason, 2014)

Tuberculosis is regarded as one of the most fearsome diseases in the world, as countries like India, China, and Indonesia are rated as the most infected (Tiwari & Maji, 2019). West African countries are not left behind as cases of tuberculosis are reported to be on the increase (Ajisafe, 2018). However, existing methods that are currently in use to mitigate the effects of tuberculosis have proven to be ineffective. Hence researchers have adopted the use of Machine learning, such as decision trees, naïve Bayes, and more to curb its effects. However, these methods are not without their limitations. The following are some works and their shortcomings:

Hassan and Khan (2017) conducted a study titled “Machine Learning based Predictive Model for Screening Mycobacterium Tuberculosis Transcriptional Regulatory Protein Inhibitors from High-Throughput Screening Dataset.” However, the techniques used in their research are resource-intensive regarding both time and space complexities. Ajisafe (2018) worked on the Bayesian Classification model in Predicting Tuberculosis Infection. However, the naïve Bayes algorithm adopted is susceptible to correlated features of datasets.
Hence, this research is motivated by the need to improve on the stated shortcomings of using support vector machines to predict the presence of tuberculosis in pulmonary chest X-rays or images.

Tuberculosis is an ancient affliction that has affected humanity throughout recorded history and prehistory. It has experienced significant outbreaks followed by periods of decline, similar to other infectious diseases, but its timeline poses challenges to conventional explanations of epidemic patterns. Mycobacterium tuberculosis is believed to have caused more deaths than any other microbial pathogen (Thomas, 2006).The genus Mycobacterium is thought to have originated over 150 million years ago. The current geographic distribution and specific ecological needs of Mycobacterium ulcerans distinguish it from its endemic regions. During the Jurassic period, these areas were among the last to be near the Gondwanaland landmass. Advances in molecular genetics and the sequencing of various M. tuberculosis strains have allowed for a more precise estimation of the emergence of mycobacteria. The low mutation rate of M. tuberculosis facilitates this estimation. According to Gutierrez and colleagues, an early ancestor of M. tuberculosis was identified in East Africa as far back as 3 million years ago, suggesting it may have infected early hominids at that time. However, it is likely that all contemporary members of the M. tuberculosis complex, including M. tuberculosis, its African variants Mycobacterium africanum and Mycobacterium canettii, as well as Mycobacterium bovis, shared a common ancestor in Africa approximately 35,000 to 15,000 years ago. Modern strains of M. tuberculosis are believed to have descended from a common ancestor around 20,000 to 15,000 years ago. Currently circulating strains are classified into six major lineages, or clades, all found in East Africa, although their global distribution varies. Analysis of the known mutation rate of M. tuberculosis suggests that much of the existing diversity among these strains emerged between 250 and 1,000 years ago (Thomas, 2006).
Thus, East Africa is considered the ancestral homeland of both tubercle bacilli and their human hosts. While archaeological evidence of any disease is generally scarce in East Africa, tuberculosis can be documented in

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.

Egypt over 5,000 years ago. Typical skeletal deformities associated with tuberculosis, including characteristic Pott's deformities, have been discovered in Egyptian mummies and are depicted in early Egyptian art. Among the early descriptions of Egyptian tuberculosis was that of A.J.E. Cave, published in 1939 in the *British Journal of Tuberculosis*. More recently, *M. tuberculosis* DNA has been amplified from tissues of Egyptian mummies, leaving no doubt as to the cause of early skeletal disease 12, 13 The written record of Egyptian tuberculosis is limited. There is no mention of it in medical papyri, although the descriptions of illnesses in these documents are not easily interpretable. However, tuberculosis is referenced in the Biblical texts of Deuteronomy and Leviticus, using the ancient Hebrew term schachepheth. By the time Europeans arrived in East Africa in the 19th century, tuberculosis was already well established in the region (Thomas, 2006).

Signs of Tuberculosis

An individual with a TB infection may not exhibit any symptoms. However, someone with active TB disease could experience any or all of the following signs: a persistent cough, ongoing fatigue, weight loss, decreased appetite, fever, coughing up blood, and night sweats. These symptoms can also be associated with other illnesses, so it is crucial to consult a healthcare professional to determine if you have TB. A person with TB disease might feel completely healthy or may only have an occasional cough. If you believe you have been exposed to TB, it is advisable to undergo a TB test.

Prediction

A prediction is a statement about what will happen in the future. It is an informed estimate, sometimes backed by evidence or data, but not always. Predictions are a part of statistical inference. While there is a specific methodology for this type of inference, predictions can be generated using various statistical techniques. In essence, statistics can be described as a means of conveying information about a sample from a population to other related populations. This is not always synonymous with making future predictions, which often pertain to specific points in time. The process of making such forecasts is referred to as forecasting. While predictions are often based on cross-sectional data, forecasting generally requires a time series approach. (Soni, 2015)

In data mining, prediction involves identifying data points based solely on the characteristics of another related data value. It is not necessarily tied to future occurrences, as the variables used may be unknown. Prediction establishes the relationship between the elements you wish to forecast for future reference. In data mining, this type of prediction is termed numeric prediction. The process includes analyzing trends, classification, pattern recognition, and relationships. By examining past events or instances, one can make predictions about future occurrences. A fortune teller makes a prediction using a crystal ball. A meteorologist uses maps and specific data to tell us about the possibility of rain, snow, or sunshine. A prediction is a statement about what you think will happen in the future. (Lobiyal, 2015).

Aim and Objectives

The goals and purposes of this research are to create tuberculosis prediction models utilizing a support vector machine and to assess the effectiveness of the developed model through precision and accuracy metrics to determine its feasibility.

Methodology

The method is categorized into three stages namely: Data collection, Pre-processing, and Classification. In data collection, pulmonary chest X-ray data was obtained from a public repository called kaggle.com. The obtained data is pre-processed using feature extraction where relevant features were extracted using the SURF feature extraction technique to improve classification performance. SURF (Speed up Robust Features) performs extraction by finding the interest points of images using equation (1.1)

$$det(H_{approx}) = D_{xx}D_{yy} - (.9D_{xy})^2 \quad (1.1)$$

where $det(H_{approx})$ represents determinant of Hessian matrix to obtain maximum value point, D_{xx} represents the convolution of the Gaussian second order derivation with image at point x, D_{yy} represents the convolution of the Gaussian second order derivation with image at y, and D_{xy} represents the convolution of the Gaussian second order derivation with image at $x\ and\ y$. Thereafter, data with extracted features will be split into training and test sets to train and validate the proposed model at the classification stage. The classifier employed for classification is the Support Vector Machine (SVM). SVM is computed as follows:

Given an unlabelled instance image$x$, SVM predicts by creating hyperplanes that linearly separate the two classes (tuberculosis and non-tuberculosis images) as shown in equation (1.2)

$$w^T x + b = 0 \quad (1.2)$$

Where w^T represents the weight vector, b is the bias factor, x is the data sample in T. The best choice of hyperplane is the hyperplane that has the maximum margin between both classes. The proposed predictive

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.

model will be implemented using Python Programming language and evaluation of the model will be carried out using accuracy, recall, f1-score, and precision.
This research is expected to establish an effective tuberculosis predictive model with high accuracy and low false positives.

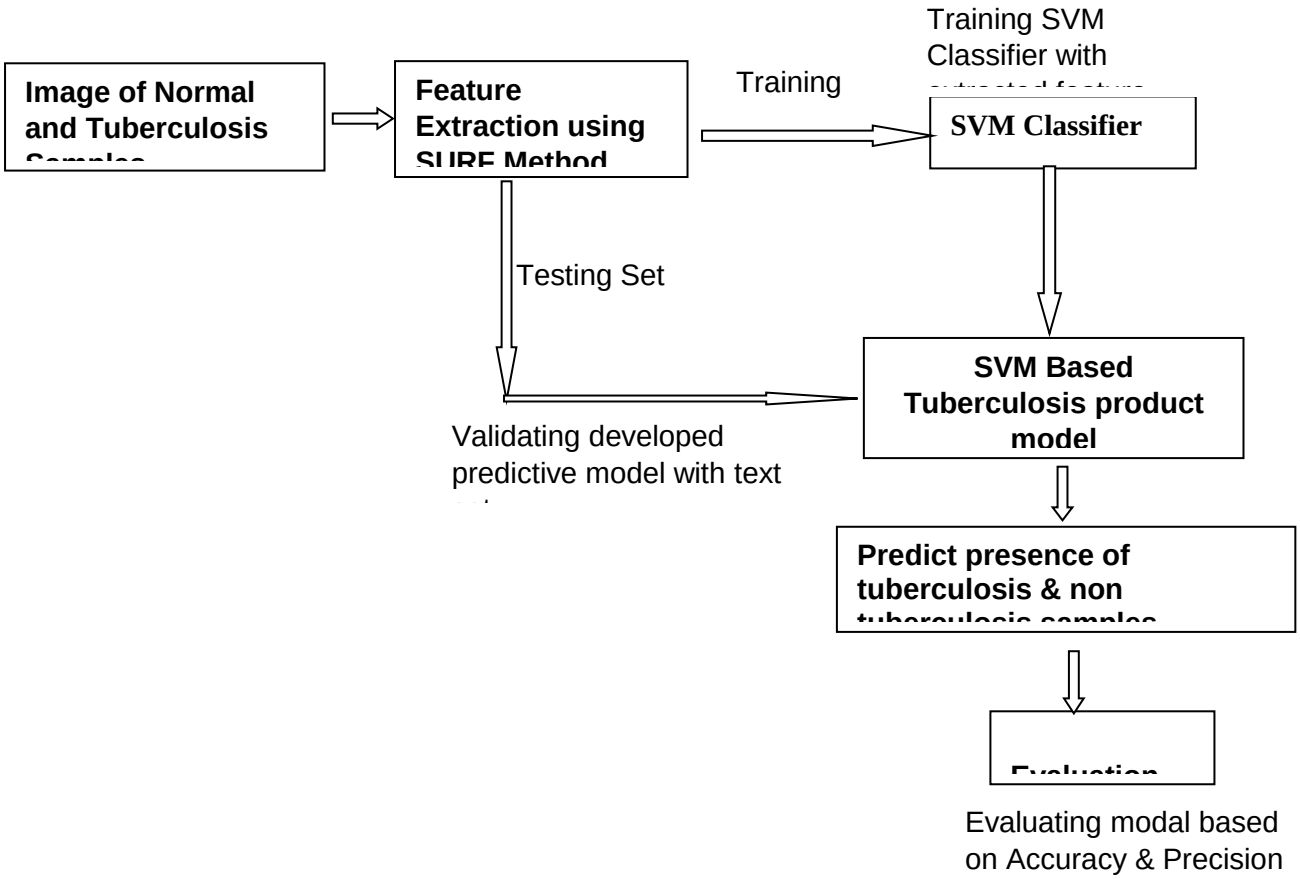

Fig. 1: System Architecture for the Prediction of Tuberculosis of Chest X-Ray Image

Data Collection

- In this research, the National Library of Medicine in Maryland, USA, collaborated with Shenzhen No. 3 People's Hospital and Guangdong Medical College in Shenzhen, China, to create a standardized digital image database for tuberculosis. Pulmonary chest X-rays were captured using Philips DR Digital Diagnose systems as part of the routine procedures in outpatient clinics. The dataset comprises 337 cases exhibiting TB symptoms and 326 normal cases, totaling 663 X-rays. Image parameters:
- Format: PNG
- Image size varies for each X-ray. It is approximately 3K x 3K.

Image file names are coded as CHNCXR_#####_0/1.png, where '0' represents the normal and '1' means the abnormal lung as shown in Table 3.1, while samples of tuberculosis and non-tuberculosis images are shown in Figure 3,2

Table 1: Class Identification

Class	Values

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.

Images with tuberculosis	CHNCXR_0001_1
Images with non-tuberculosis	CHNCXR_0004_0

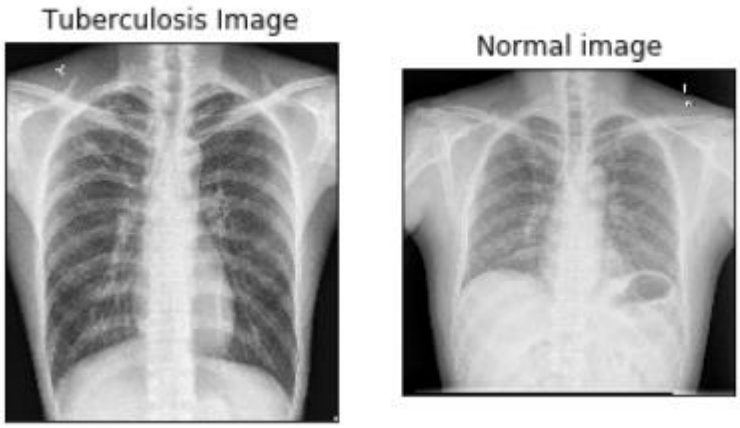

Figure 2**:** Sample of tuberculosis and non-tuberculosis images

Surf Feature Extraction On Pulmonary Chest X-Ray

A local feature descriptor called SURF was created to address the drawbacks of SIFT (Scale Invariant Features Transform). For improved classification performance and lower time costs, it is quick and precise in identifying important aspects of photos. Surf employs the sun of Haar wavelength responses for orientation assignment and the Hessian matrix approximation to find interesting locations. For example, the hessian matrix H(x,σ) in x at scale σ is provided as follows given a point (x,y) in an image:

$$H(x,\sigma) = \begin{bmatrix} Lxx(x,\sigma) & Lxy(x,\sigma) \\ Lxy(x,\sigma) & Lyy(x,\sigma) \end{bmatrix} \quad (1)$$

Where $Lxx(x,\sigma)$ represents a convolution of the gaussian second order derivative $\frac{\partial^2}{\partial x^2}$with image I at point x, and similarly for $Lxy(x,\sigma)$ and $Lyy(x,\sigma)$ as the result is given as:

$$H(x,y) = \begin{bmatrix} \frac{\partial^2}{\partial x^2} & \frac{\partial^2}{\partial x \partial y} \\ \frac{\partial^2}{\partial x \partial y} & \frac{\partial^2}{\partial y^2} \end{bmatrix} \quad (.2)$$

$H((x,y))$ Is computed to obtain the extreme points, which are subsequently taken as surf feature points with the approximate second-order derivation of Gaussian denoted as:

$$det(H_{approx}) = D_{xx}D_{yy} - (.9D_{xy})^2 \quad (3)$$

where $det(H_{approx})$ represents determinant of Hessian matrix to obtain maximum value point, D_{xx} represents the convolution of the Gaussian second order derivation with image at point x, D_{yy} represents the convolution of the Gaussian second order derivation with image at y, and D_{xy} represents the convolution of the Gaussian second order derivation with image at x and y

Evaluation

Once tuberculosis and non-tuberculosis have been classified, test data must be used to determine the quality of the established model. Specifically, the model's accuracy in differentiating between samples with and without tuberculosis should be estimated. This was accomplished by employing the confusion matrix to determine metrics like True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN).

Classification

Before classification, images with extracted features were split into training and test sets. However, in this research, a fraction of the dataset was used as the data. Hence, SVM was trained iteratively with 200 samples while 100 samples were used to validate the developed model.

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.

Table 2 shows the training and test samples used for the classification process.

Class Label	Training Set	Test Set
Tuberculosis	87	50
Non-tuberculosis	113	50
Total	**200**	**100**

Result

The evaluation and results of the created tuberculosis prediction system's testing using the test set are presented in this part. The confusion matrix result of the generated model's prediction on test data is displayed, The created model's accuracy, recall, and precision are displayed in Table 2

Table 3: Confusion Matrix Result of the Developed Tuberculosis Predictive Model

		Predicted Class	
		Non-tuberculosis	Tuberculosis
Actual Class	Non-tuberculosis	$TN = 47$	$FP = 4$
	Tuberculosis	$FN = 3$	$TP = 46$

Table 3 presents the results of the Confusion Matrix for the tuberculosis prediction model when tested on the dataset. The findings indicate that, from 100 test samples, the model accurately identified 46 out of 50 actual tuberculosis cases as positive, while incorrectly classifying 4 tuberculosis cases as non-tuberculosis. Additionally, among the 50 actual non-tuberculosis samples, the model correctly classified 47 as non-tuberculosis, but mistakenly identified 3 non-tuberculosis cases as tuberculosis. The mathematical representation is as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{46+47}{46+47+4+3} = \frac{93}{100} = 0.93 * 100 = \mathbf{93.0}\%$$

$$Precision = \frac{TP}{TP+FP} * 100 = \frac{46}{46+4} = \frac{46}{50} = 0.92 * 100 = \mathbf{92.0}\%$$

$$False\ Alarm\ Rate = \frac{FP}{TN+FP} = \frac{4}{47+4} = 0.078$$

$$Recall = \frac{TP}{TP+FN} = \frac{46}{46+3} = 0{,}94 * 100 = \mathbf{94.0}\%$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{0.92 * 0.94}{0.92 + 0.94} = 0.9298 * 100 = \mathbf{92.99}\%$$

Table 4: Number of correct and incorrect classifications obtained by the developed tuberculosis prediction model.

Total number of test instances	Correctly Classified $(TN + TP)$	Incorrectly Classified $(FP + FN)$
100	**93**	**7**

Table 4: shows the total number of correct and incorrect classifications obtained after the developed tuberculosis predictive model was validated with the test data. The number of correct and incorrect classifications was computed by summing the number of True Positives (TP) and True Negatives (TN), and some False Negatives (FN) and False Positives (FP) obtained by the developed prostate cancer predictive model in Table 2 as follows:

Correct Classification $= TP + TN$

$= 46 + 47 = \mathbf{93}$

Incorrect Classification $= FN + FP$

$= 4 + 3 = \mathbf{7}$

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.

Table 5: Evaluation of the Developed model using Accuracy, Precision, Recall, and F1-Score

Accuracy	Precision	Recall	F1-Score
%	%	%	%
93.0	92.0	94.0	92.99

Based on the derived values in Table 5, the developed model is revealed to be effective in predicting both non-tuberculosis and tuberculosis instances with Accuracy of 0.93 (93.0%), Precision of 0.92 (92.0%), Recall of 0.94 (94.0%), F1-Score of 0.9299 (92.99%). The general performance of the developed model on the test data was also impressive as it attained a low false alarm rate of 0.078.

Conclusion

A tuberculosis prediction model based on support vector machines was developed to achieve the objectives of this study. The pulmonary chest X-ray dataset, sourced from kaggle.com, served as the benchmark dataset to fulfil these objectives. The dataset underwent pre-processing using the SURF feature extraction technique to identify key points in the images, and it was subsequently divided into training and testing sets. Approximately 500 significant feature points were extracted and utilized for training and testing the support vector classifier in order to construct the proposed prediction model. The models were assessed using standard performance metrics, including accuracy, recall, precision, F1-score, and false alarm rate. The results demonstrate the effectiveness of the proposed model, achieving an accuracy of 93.0%, a precision of 92.0%, a recall of 94.0%, an F1-score of 92.99%, and a false alarm rate of 0.078.

References

Ajisafe, B. (2018). Bayesian classification for tuberculosis prediction. *African Journal of Infectious Diseases*, 14(7), 405-412.

Alemu, B. (2018). Tuberculosis and its spread. *Journal of Medical Microbiology*, 54(2), 89-95.

Davidson, S. (2009). The use of data mining techniques for tuberculosis diagnosis. *International Journal of Health Informatics*, 15(4), 120-128.

Hassan, M., & Khan, R. (2017). Machine learning model for screening Mycobacterium tuberculosis transcriptional regulatory protein inhibitors. *Journal of Biomedical Research*, 22(9), 511-522.

Jason, B. (2014). Feature selection in machine learning. *Machine Learning Advances*, 27(5), 155-172.

Lobiyal, D. K. (2015). Predictive methods in data mining. *Data Mining and Knowledge Discovery*, 23(2), 200-215.

Soni, A. K. (2015). Prediction techniques in data mining. *Statistics in Health Science*, 11(3), 210-230.

Tamer, B. J (2011). Understanding the basics of data mining. *Computer Science Review*, 10(3), 240-260.

Thomas, D. K (2006). The history of tuberculosis and its evolution. *Journal of Infectious Diseases*, 31(7), 98-110.

Tiwari, A., & Maji, S. (2019). Tuberculosis incidence in high-risk countries. *WHO Global Health Reports*, 34(6), 340-352.

Cite this article as:
Olojido, J.B.,& Oladimeji, B. J. (2025). Prediction of tuberculosis using machine learning techniques. *FNAS Journal of Mathematical and Statistical Computing, 2*(2), 8-14.