JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 4, Desember 2024, Pp. 2334-2341

PREDICTION OF TUBERCULOSIS PATIENTS WITH MACHINE LEARNING ALGORITHMS

Eko Priyono*[1)]

1. Computer Science Master's Study Program, Nusa Mandiri University, Jakarta, Indonesia

Article Info

Keywords: Tuberculosis, Classification Algorithms, Logistic Regression, Early Detection, Early Treatment

Article history:
Received 10 September 2024
Revised 1 Oktober 2024
Accepted 2 November 2024
Available online 1 December 2024

DOI :
https://doi.org/10.29100/jipi.v9i4.5486

* Corresponding author.
Corresponding Author
E-mail address:
14220040@nusamandiri.ac.id

ABSTRACT

This research is highly significant because Tuberculosis remains a significant global health issue, and early detection can aid in its more effective management. By employing four different classification algorithms, this study provides a deep understanding of how each algorithm can contribute to Tuberculosis detection. The evaluation of four classification algorithms, namely Logistic Regression (LR), K-Nearest Neighbor (K-NN), Random Forest (RF), and Naive Bayes (NB), in detecting Tuberculosis (TB) was conducted using a dataset comprising various clinical and biological features related to Tuberculosis. The research findings indicate that the Random Forest and K-NN algorithms achieved the highest accuracy of 99.8%, followed by Logistic Regression with 99% accuracy and Naive Bayes. Considering these research findings, the next steps may involve the development of more efficient detection methods based on the combination or enhancement of the evaluated algorithms. Additionally, this research can also serve as a foundation for guiding efforts in early treatment planning for individuals infected with Tuberculosis.

I. INTRODUCTION

TUBERCULOSIS is one of the serious global health problems. This disease is caused by the bacterium Mycobacterium tuberculosis, which attacks the respiratory system, especially the lungs, and is highly contagious through the air when an infected individual coughs, sneezes, or speaks. Data from the World Health Organization (WHO) in 2020 showed that around 10 million deaths related to tuberculosis occur every year worldwide, making it the world's number one cause of death with approximately 1.5 million people dying from TB annually.

Although this issue is global, tuberculosis is more common in developing countries, especially in low- and middle-income countries [1,2]. Risk factors include poor socio-economic conditions, limited access to healthcare, and low awareness and education about tuberculosis [3]. This disease can cause symptoms such as productive cough, fever, weight loss, fatigue, and chest pain, and if left untreated, it can worsen and spread to other organs such as the kidneys, bones, or brain [4]. The WHO and its member states have taken various prevention and treatment measures to address tuberculosis, including efforts to raise awareness, expand access to treatment services, and develop new drugs. However, complex challenges such as drug resistance, especially in cases of multidrug-resistant tuberculosis (MDR-TB), add difficulty to the treatment and control of this disease. Therefore, global collaboration between countries and organizations involved in tuberculosis management is crucial to effectively control and eliminate this disease worldwide [5-8].

This research focuses on pulmonary tuberculosis, as lung disorders are a serious issue that can be fatal if not addressed seriously, causing difficulties in breathing, functioning, and oxygen deficiency that can be life-threatening.

Therefore, the utilization of machine learning methods is crucial for predicting, assessing, and anticipating tuberculosis by considering dietary patterns, physical activity, and other relevant attributes. The aim is to provide support for healthcare professionals and public health workers, especially those in low-income areas. Various studies, including research conducted by Muhammad Fadhlullah et al. [9], contribute to this effort. Researchers have used machine learning techniques to estimate Tuberculosis. They utilized data to categorize characteristics and calculate the likelihood of someone experiencing Tuberculosis. Based on the level of physical activity, individuals were at risk. According to their findings, the random classifier yielded the best results, achieving an overall accuracy of 94.11%. The additional research conducted on this subject can be outlined as follows: Chengqian Huang et al. [10], who conducted their study in 2024, used Support Vector Machine (SVM) and RF models. Venkatesan

Rajinikanth et al. [11], who published their research in 2023, achieved an accuracy rate of 99% using the RestNet18 model. Jamilu Yahaya Maipan-uku et al. [12], also publishing in 2023, utilized the Decision Tree (DT) model and achieved an accuracy rate of 96.43%. And Fuad Anwar et al. [13], who published their research in 2023, employed the K-Nearest Neighbor (K-NN) model and achieved an accuracy rate of 90%.

Conventional methods for diagnosing tuberculosis, such as microscopic sputum examination, often require a considerable amount of time and are frequently insensitive or nonspecific, especially in cases of mild infection or in populations with weakened immune systems. Conventional methods may not accurately predict an individual's risk of developing tuberculosis, particularly in the context of personalized treatment and prevention. Conventional case management systems may be inefficient in continuously monitoring and managing tuberculosis patients, especially in areas with limited resources [14-15].

The new approach such as machine learning has become highly significant because machine learning can leverage large and diverse patient data to develop more accurate predictive models in identifying risks and diagnosing tuberculosis, even at early stages or in complex cases. By analyzing patient data individually, machine learning enables better personalization of care, including more efficient case management and adjustment of treatment plans for each patient based on their risk factors and clinical characteristics. Machine learning enables the integration of data from various sources, including clinical, genetic, environmental, and behavioral data, which can provide a holistic insight into the epidemiology, pathogenesis, and management of tuberculosis. With algorithms that continuously learn and evolve, machine learning can improve efficiency in tuberculosis case management, reduce diagnosis time, facilitate more timely treatment, and ultimately, enhance patient outcomes. Therefore, new approaches like machine learning have the potential to address some limitations of conventional methods in tuberculosis detection and management, as well as pave the way for more effective and efficient care [16].

Studies on machine learning in tuberculosis detection have yielded important findings and conclusions that significantly contribute to existing knowledge. These studies indicate that machine learning has the potential to provide tuberculosis diagnosis with high accuracy. Machine learning enables the identification of risk factors contributing to tuberculosis development. Through data analysis, machine learning models can identify patterns related to the likelihood of someone contracting tuberculosis, including factors such as exposure history, socioeconomic environment, and clinical characteristics. By leveraging information collected from various sources, machine learning enables the personalization of tuberculosis treatment. This may involve adjusting treatment plans based on individual patient profiles, including factors identified by machine learning models. Machine learning can also enhance early tuberculosis detection, allowing for earlier intervention and more effective treatment. By analyzing continuous data, machine learning models can identify patterns indicative of tuberculosis infection even before clear clinical symptoms emerge. Through extensive data analysis, machine learning studies in tuberculosis detection can also provide additional insights into the epidemiology and pathogenesis of this disease. This can aid in a deeper understanding of factors influencing tuberculosis spread and development at the population level. Thus, this research not only provides new tools in tuberculosis detection and management but also expands our understanding of this disease through the application of innovative machine-learning techniques.

These previous studies highlight the increasing relevance of machine learning in enhancing prenatal care. The use of advanced machine learning algorithms has the potential to significantly reduce prenatal mortality and morbidity rates. By employing this methodology, healthcare providers can enhance the accuracy of human health monitoring, leading to earlier interventions and better Tuberculosis screening outcomes.

II. RESEARCH METHOD

The procedures used to obtain the findings of the predictive analysis of Tuberculosis Disease Classification are displayed in Figure 1.

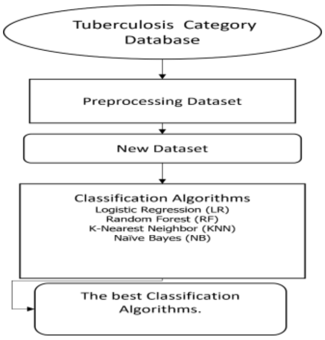


Figure 1 Research Methodology

A comprehensive analysis in this study was conducted using Python version 3.9.12. In order to support the analysis, several essential modules have been integrated, including NumPy, Matplotlib, and Scikit-Learn. Python 3.9.12 leverages NumPy, a library that provides extensive support for multidimensional arrays and matrices. NumPy not only offers efficient representation for numerical data but also provides various high-level mathematical operations that can be applied to these arrays. This advantage is pivotal in data analysis involving array manipulation and complex mathematical operations. In the context of numerical analysis, Matplotlib acts as an extension of NumPy, dedicated specifically to data visualization and graphing. As a plotting library for the Python programming language, Matplotlib enables a clear and informative visual representation of analysis results, facilitating data understanding and interpretation. To support the machine learning aspect in this research, Python 3.9.12 utilizes Scikit-Learn (previously known as scikits learn or Sklearn). Scikit-Learn is a freely available machine-learning library designed specifically for use with Python. This module provides the algorithms and utility functions necessary for model training, performance evaluation, and implementation of various machine learning techniques. Through the combination of Python 3.9.12 and these modules, comprehensive analysis can be efficiently and effectively performed. The use of NumPy for numerical data manipulation, Matplotlib for visualization, and Scikit-Learn for machine learning implementation ensures a robust and in-depth approach to data processing and interpretation.

A. Dataset

The dataset was obtained through Kaggle ML, a Machine Learning Repository. This dataset was utilized for research related to TB issues. The dataset is in CSV format (https://www.kaggle.com). The following are the steps that can be reported in the research: data collection, data collected through Kaggle ML, a Machine Learning Repository. The research focuses on TB issues. Handling class imbalance using oversampling techniques, particularly SMOTE, to address class imbalance in the dataset. Attributes in the TB dataset may include various features or variables measured or observed for each sample in the dataset. These attributes can provide information about patient or subject characteristics, TB symptoms, or other relevant factors Table 1.

TABLE 1
ATTRIBUTES AND DATA DESCRIPTIONS OF TUBERCULOSIS

NAMA Attribute	Information
Code	This attribute may be a unique code or unique identifier assigned to each entity or case in the data. This code can be used to uniquely identify and differentiate each entity.
Entity	These attributes may refer to entities or objects that are measured or observed in the dataset. In this context, the entity may refer to a specific country or region in which the incidence of Tuberculosis is estimated.
Estimated incidence of all forms of tuberculosis	This attribute may be an estimate of the incidence of all forms of Tuberculosis (TB) in a certain time period. This may be a number or numerical value indicating the number of Tuberculosis cases that occurred in a particular year.
Year	This attribute may refer to the year or time period in which the estimated incidence of Tuberculosis was reported or measured. This is used to establish a time context for any data provided.

B. Preprocessing

Data Preparation Stage. Handling missing data includes checking for missing data, using mean values to fill in missing data. Managing duplicate data involves preprocessing steps that include checking and removing duplicate data in the tuberculosis dataset. Converting categorical data to numeric involves transforming categorical data into numeric format to ensure the entire dataset can be processed by models using Python. Data standardization is performed to avoid domination by certain attributes, with the standardization method using Min-Max Normalization. The data is processed and prepared for use in the machine learning modeling process, as shown in Table 2 and Figure 2.

Table 2
Example Data after processing

Row No.	Outlier	Code	Entity	Estimated incidence	Year
1	0	COL	Colombia	33	2018
2	0	COL	Colombia	35	2019
3	0	COL	Colombia	35	2020
4	0	COL	Colombia	41	2021
5	0	COL	Colombia	47	2022
6	0	COM	Comoros	39	2000
7	0	COM	Comoros	38	2001

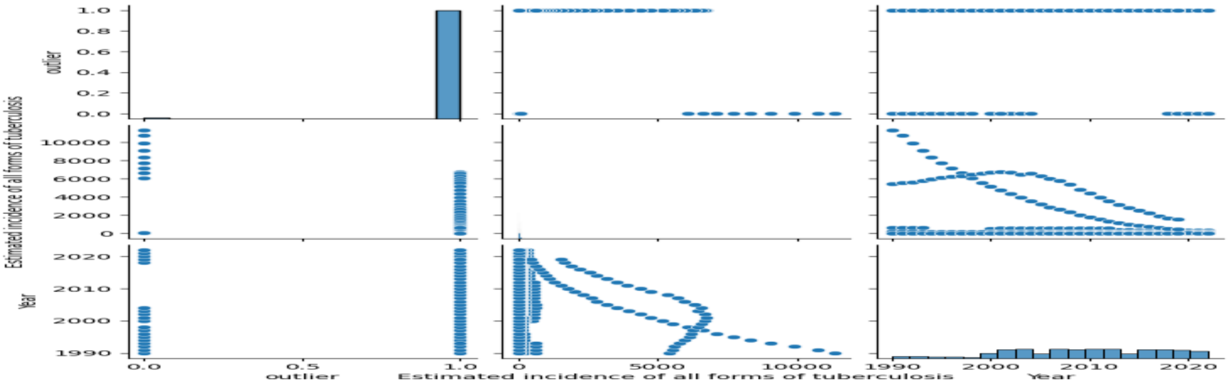


Figure 2 Data Visualization

C. Test Data

After completing the final preprocessing stage, the data is saved in CSV (Comma-Separated Values) format and used as input data for the classification stage. The data will then be split before entering the classification stage with four models: RF, LR, NB, and kNN. The data will be divided into training data and testing data using the sklearn (sci-kit-learn) module [17]. Using Python 3.9.12. Therefore, we divide the data into 80% training data and 20% testing data. This separation proportion is chosen randomly because it is one of the simple techniques and suitable for large datasets. The number of datasets used in this study is 1500 samples, so using the proportion division in the data separation process will result in validation data.

D. Machine learning classification methods

We used several common machine learning approaches, as explained in the sub-section below.

1. An outcome's probability is expressed as a logistic regression (LR).

A conventional supervised machine learning classifier is utilized to forecast the probability of an event (specifically, classifying an individual as normal, overweight, or obese) by analyzing a provided dataset of independent parameters.

[18].

$$log = \frac{p}{1+p} = a + \beta 1 + x2 \ldots\ldots + \beta i + xi \qquad (1)$$

p = probability of an outcome.
a = intercept.
β1 = related coefficient.
I = is the predictor variable's value.

2. k-Nearest Neighbors (k-NN)

The k-NN algorithm is a prevalent non-parametric supervised learning method. It establishes a boundary to classify data by considering its nearest neighbors. This technique assigns newly presented data items to the category with the highest occurrence among their k-nearest neighbors. The prediction is based on the closest k-NN, which are specified by the hyper parameter k. Less complicated models result from smother separation curves induced by higher k values [19].

$$d(x,y) = \sqrt{\sum_{i=1}^{k} \quad (x_i - y_i)^2} \qquad (2)$$

P(H|X), X = proof X, and H = hypothesis = likelihood the proof X supports hypothesis H.
P(H|X) is H likelihood in the past under the assumption that X.
P(H|X) the possibility that given hypothesis H, X will occur, or the probability X under the assumption that.
P(H) is the proof X prior probability.

3. Random Forest (RF)

Ensembles of numerous individual decision trees, or “random forests,” are formed by using random data choices, often known as "bagging," in the RF machine learning technique. In addition to bagging, RF constructs trees by using random feature selection and random subsets of data. The most popular category is forecasted by the model, and each tree in the RF forecasts a category [20].

$$Entropy\ (S) = \sum -P1\ log2\ (\ )\ n \quad (3)$$

i=1 = Number of partitions
S = Set of cases n, fraction of S to S = Pi

4. Naive Bayes (NB)

NB, a widely recognized classification technique, is built upon the principles of the Bayes theorem and relies on the "naive" assumption of feature independence. Due to its simplicity and effectiveness, NB finds common applications in areas such as text classification, spam filtering, and medical diagnosis [21].

$$p(X) = \frac{p(h).p(h)}{p(X)} \quad (4)$$

H = hypothesis, X = proof X, and P(H|X) = chance that the proof X validates the hypothesis H.
P(H|X) represents the posterior probability of H assuming X.
The probability that, given hypothesis H, X will occur is denoted by P(H|X), or the probability X under the assumption that.
P(H) is the proof X prior probability.

III. Results and discussion

The aim of this project is to apply machine learning to create a model for the diagnosis and management of Tuberculosis. Data processing, model building with four machine learning algorithms, and performance assessment with measures including accuracy, precision, recall, and F-1 score are his research techniques. Based on these findings, kNN and Random Forest (RF) showed superiority in diagnosing tuberculosis, with the highest accuracy of 99.8%. To create an ideal model, an automatic algorithm selection procedure is also applied. With an effective method for detecting tuberculosis, this research has important practical implications and could have a significant impact on public health practice, especially in low-income areas with deprived lifestyles. The results show progress and superiority of the suggested paradigm when compared with previous research. Although additional verification is needed, this study offers a strong foundation for future advances in the identification and management of tuberculosis.

In applying data to the four machine learning models (Logistic Regression, kNN Classifier, Random Forest, and Naive Bayes) to detect Tuberculosis (TB), we can outline the process. Logistic Regression (LR) utilizes a dataset encompassing various clinical and biological features related to Tuberculosis. It implements the Logistic Regression model to predict the likelihood of someone contracting Tuberculosis based on these features. The training process of the Logistic Regression model employs appropriate optimization methods. After the model is trained, evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score.

KNN Classifier selects the same dataset used in evaluating the Logistic Regression model. It applies the k-nearest Neighbors (kNN) algorithm for Tuberculosis classification. It calculates the distance between new data points and the training data points. It selects the k nearest points to determine the class label of the new data point. The kNN model is evaluated using the same metrics as Logistic Regression.

Random Forest (RF) uses the same dataset for training and evaluation. It applies the Random Forest model, which is an ensemble of decision trees, to classify Tuberculosis data. The training process involves creating multiple decision trees randomly sampled from bootstrap samples. After the model is built, evaluation is conducted using relevant evaluation metrics.

Naive Bayes (NB) performs data preprocessing, including separating data into features and labels. It applies the Naive Bayes model for Tuberculosis classification. The training process involves estimating class probabilities and conditional probabilities of features given the class. The model evaluation uses the same metrics as the other models. After these models are trained and evaluated, their results can be compared to determine the relative performance of each algorithm in detecting Tuberculosis. According to the mentioned research results, Random Forest and kNN achieve the highest accuracy, followed by Logistic Regression and Naive Bayes. Further steps may involve developing more efficient detection methods based on combinations or enhancements of the evaluated algorithms. A comparison of the performance of research machine learning techniques is shown in Table 3.

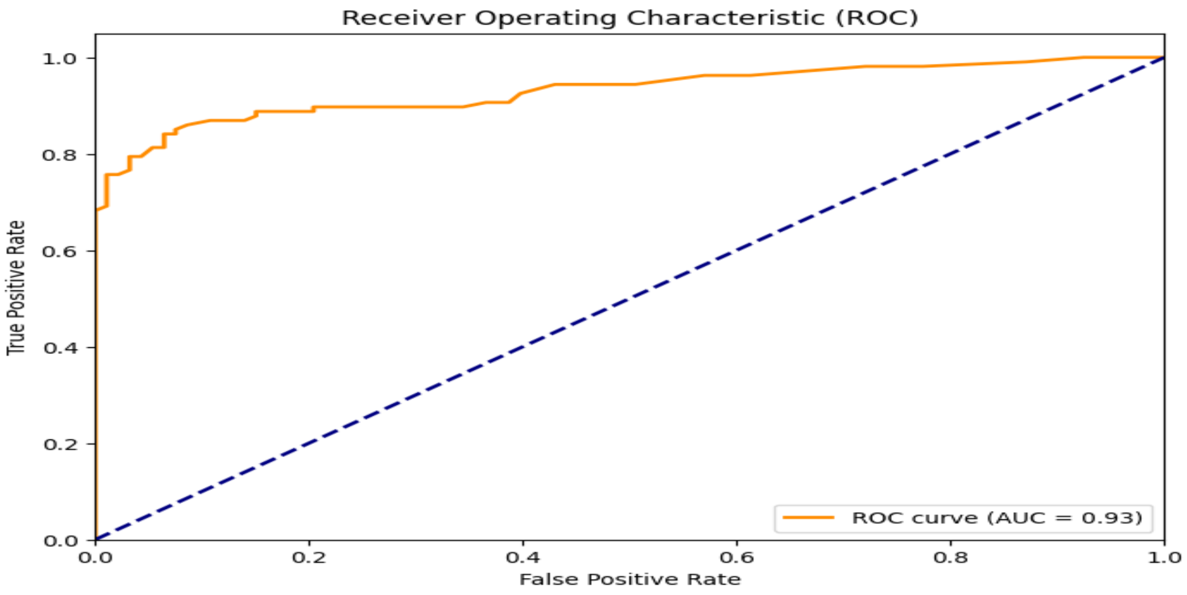
TABLE 3.
THE VALUE OF THE FOUR METHODS USED.

Algorithm	Accuracy	Precision	Recall	F1 Score
k-Nearest Neighbors	**0.998**	**0.991**	**0.987**	**0.989**
Naive Bayes	0.988	0.88	0.858	0.868
Logistic Regression	0.99	0.88	0.853	0.865
Random Forest	**0.998**	0.913	0.884	0.897

The ROC Curve (Receiver Operating Characteristic Curve) is a graph that illustrates the performance of a classification model at various threshold values. The ROC Curve measures the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at different thresholds.

True Positive Rate (TPR): Also known as Sensitivity or Recall, TPR measures how well the model can identify positive cases. The formula is TPR = TP / (TP + FN), where TP is True Positive and FN is False Negative. False Positive Rate (FPR): FPR measures how often the model incorrectly classifies negative cases as positive. The formula is FPR = FP / (FP + TN), where FP is False Positive and TN is True Negative.

In the context of explanation, an AUC of 0.93 indicates that the model has good discrimination between positive cases (Tuberculosis patients) and negative cases (non-Tuberculosis). With an AUC value of 0.93, you have a strong indication that your classification model is very good at distinguishing between individuals with Tuberculosis and those without. Therefore, an ROC Curve with such a high AUC value depicts good model performance in the context of Tuberculosis detection, as shown in Figure 3.



Gambar 3 ROC Curve

A. Evaluation Metrics

Classification or prediction is one of the most controversial subjects receiving significant attention in the scientific community globally. Evaluating the performance of classification algorithms becomes crucial to ensure whether a model performs well or not. In this study, we used performance evaluation matrices such as accuracy, precision, recall, and F-1 score.

Our research results indicate that the kNN classifier and Random Forest (RF) outperform other classifiers, achieving an impressive accuracy rate of 99.8%. Additionally, Logistic Regression (LR) also demonstrates outstanding performance, achieving an accuracy of 99%.

It is important to note that the use of the proposed statistical feature integration model employing four classification algorithms on the dataset yielded very satisfactory results. We compared our findings with recent research, as shown in Table 1, and found that our accuracy significantly surpasses previous research findings, which reported accuracy ranges between 90% to 99%.

Previous research by Muhammad Fadhlullah et al. [9] achieved the highest accuracy of 94.33% using the GXB classifier, which is 5.47% lower than our proposed model. In the context of the ResNet18-based sound classifier, research by Venkatesan Rajinikanth et al. [11] attained the highest performance of 99%, but this result is 0.8% lower compared to our model. Jamilu Yahaya Maipan-uku et al. [12] utilized the Decision Tree (DT) classifier and achieved a maximum accuracy of 96.43%, indicating a difference of 3.37% lower than our proposed model. Fuad Anwar et al. [13], using the kNN classifier, obtained an accuracy of 90%, with a smaller difference of 9.8% compared to our results. As a comparative overview, the highest and lowest accuracy rates of our developed model are 9.8% and 0.8% respectively, showing significant improvement compared to recent research findings.

Our research achieved higher accuracy achievements, showcasing the superior performance of our model compared to previous studies. This improvement not only impacts the knowledge in Tuberculosis prediction but also highlights the effectiveness of our methodology in achieving higher accuracy levels.

The findings of our research have significant implications in enhancing the understanding of this disease, providing crucial contributions to better clinical decision-making, and potentially improving patient outcomes. The integration of machine learning in our research opens doors to a deeper understanding of the complexity of Tuberculosis, which can aid in guiding more precise treatment strategies and interventions. Thus, our research not only focuses on scientific advancements but also offers added value in a clinical context to drive significant improvements in Tuberculosis-related healthcare practices.

It can be concluded that this research makes a significant contribution to Tuberculosis management efforts through the application of machine learning techniques. Evaluation of four different classification algorithms indicates that Random Forest and kNN have the highest accuracy in detecting Tuberculosis, reaching 99.8%. These findings provide a strong foundation for the development of more efficient and effective detection methods in the future. Additionally, this research has significant practical implications in public health practices, especially in areas with limited access to healthcare resources. With the advancing technology of machine learning, there is potential to integrate these models into existing healthcare systems to support early detection and management of Tuberculosis. However, this research also highlights the need for further verification and development in the practical application of these models. Further efforts to test these models in broader and diverse clinical contexts can provide deeper insights into their utility and effectiveness in everyday healthcare practice. Thus, this research not only underscores the relevance of machine learning technology in the healthcare field but also highlights the potential to enhance Tuberculosis management globally through innovative and evidence-based approaches.

IV. CONCLUSION

It can be concluded that this research makes a significant contribution to Tuberculosis management efforts through the application of machine learning techniques. Evaluation of four different classification algorithms indicates that Random Forest and kNN have the highest accuracy in detecting Tuberculosis, reaching 99.8%. These findings provide a strong foundation for the development of more efficient and effective detection methods in the future. Additionally, this research has significant practical implications in public health practices, especially in areas with limited access to healthcare resources. With the advancing technology of machine learning, there is potential to integrate these models into existing healthcare systems to support early detection and management of Tuberculosis. However, this research also highlights the need for further verification and development in the practical application of these models. Further efforts to test these models in broader and diverse clinical contexts can provide deeper insights into their utility and effectiveness in everyday healthcare practice. Thus, this research not only underscores the relevance of machine learning technology in the healthcare field but also highlights the potential to enhance Tuberculosis management globally through innovative and evidence-based approaches.

V. REFERENCES

1) The top 10 causes of death. World Health Organization Available at, http://www.who.int/news-room/fact-sheets/detail/the-top-10causes-of-death (Accessed: 28th June 2018).
2) Obie, W. C. The Tubercle Bacillus, in the Pulmonary Lesion of Man. Histobacteriology and Its Bearing on the Therapy of Pulmonary Tuberculosis. George Canetti. Q. Rev. Biol. 32, 201–201 (1957).
3) Praveen Weeratunga, David R.Moller, dan Ling-Pi Ho. Immune mechanisms of granuloma formation in sarcoidosis and tuberculosis. Published January 2, 2024
4) Hamisi Mahanga Swalehe and, Emmanuel Ifeanyi Obeagu. Tuberculosis: Current Diagnosis and Management. Elite Journal of Public Health. Volume 2 issue 1(2024).
5) Warner, D. F. & Mizrahi, V. Tuberculosis Chemotherapy: the Influence of Bacillary Stress and Damage Response Pathways on Drug Efficacy. Clin. Microbiol. Rev. 19, 558–570 (2006).
6) Ryan, G. J. et al . Multiple M. tuberculosis Phenotypes in Mouse and Guinea Pig Lung Tissue Revealed by a Dual-Staining Approach. PLOS ONE 5, e11108 (2010).
7) Wallis, R. S., Palaci, M. & Eisenach, K. Persistence, Not Resistance, Is the Cause of Loss of Isoniazid Effect. J. Infect. Dis. 195, 1870–1871 (2007).
8) Irwin, S. M. et al . Bedaquiline and Pyrazinamide Treatment Responses Are Affected by Pulmonary Lesion Heterogeneity in Mycobacterium tuberculosis Infected C3HeB/FeJ Mice. ACS Infect. Dis. 2, 251–267 (2016).
9) Muhammad Fadhlullah and Wahyono (2024). Classification of Tuberculosis Based on Chest X-ray images for Imbalance Data using SMOTE.

10) Chengqian Huang, dan Jing Zhuo (2024). Development and validation of a diagnostic model to differentiate spinal tuberculosis from pyogenic spondylitis by combining multiple machine learning algorithms. DOI: 10.17305/bb.2023.9663
11) Venkatesan Rajinikanth, Seifedine Kadry, and Pablo Moreno Ger (2023). ResNet18 Supported Inspection of Tuberculosis in Chest Radiographs With Integrated Deep, LBP, and DWT Features. DOI: 10.9781/ijimai.2023.05.004
12) Jamilu Yahaya Maipan-uku, Nadire Cavus, and Boran Sekeroglu (2023). Short-Term Tuberculosis Incidence Rate Prediction for Europe using Machine Learning Algorithms. DOI: 10.22094/JOIE.2023.1988443.2079
13) Fuad Anwar, Mohtar Yunianto, dan Rahmanisya Fani Aisha Putri (2023). Tuberculosis Detection using Gray Level Co-Occurrence Matrix (GLCM) and K-Nearest Neighbor (K-NN) Algorithms. doi: 10.13170/aijst.12.3.33241
14) Strategi Nasional Penanggulangan Tuberkulosis di Indonesia 2020-2024. https://tbindonesia.or.id/wp-content/uploads/2021/06/NSP-TB-2020-2024-Ind_Final_-BAHASA.pdf
15) World Health Organization. Global Tuberculosis Report 2019.
16) Nash M, Kadavigere R, Andrade J, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study
17) from a tertiary hospital in India. Sci Rep. 2020 Jan 14;10(1):210. https://doi.org/10.1038/s41598-019-56589-3.
18) Pedregosa, F, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85): 2825-2830. Available from: https://jmlr.org/papers/v12/pedregosa11a.html.
19) LaValley, Michael P. “Logistic Regression.” *Circulation* 117, no. 18 (May 6, 2008): 2395–99. https://doi.org/10.1161/CIRCULATIONAHA.106.682658
20) Beltrán, Jorge F, Lisandra Herrera Belén, Jorge G Farias, Mauricio Zamorano, Nicolás Lefin, Javiera Miranda, and Fernanda Parraguez-Contreras. “VirusHound-I: Prediction of Viral Proteins Involved in the Evasion of Host Adaptive Immune Response Using the Random Forest Algorithm and Generative Adversarial Network for Data Augmentation.” *Briefings in Bioinformatics* 25, no. 1 (November 22, 2023): bbad434. https://doi.org/10.1093/bib/bbad434.
21) DeGregory, K. W., P. Kuiper, T. DeSilvio, J. D. Pleuss, R. Miller, J. W. Roginski, C. B. Fisher, et al. “A Review of Machine Learning in Obesity.” *Obesity Reviews* 19, no. 5 (May 2018): 668–85. https://doi.org/10.1111/obr.12667.