International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Sustainable Healthcare: Machine Learning-Based Tuberculosis Detection Using C4.5 Decision Tree

Helanmary M Sunny*[1], Dr.Anju Pratap[2], Christina Thankam Sajan[3], Sandhra Merin Sabu[4]

Submitted: 12/01/2025 **Revised:** 25/02/2025 **Accepted:** 08/03/2025

Abstract: Tuberculosis (TB) remains a global health crisis, particularly in resource-limited regions where diagnostic infrastructure is scarce. While deep learning models dominate recent research, classical machine learning (ML) methods offer interpretability and computational efficiency—critical for low-resource settings. This study presents the first systematic comparison of 13 ML algorithms, including C4.5 decision trees, logistic regression, and ensemble methods, for TB detection using the Shenzhen chest X-ray dataset .The C4.5 decision tree achieved near perfect accuracy (99.78%) and the lowest training time (0.147s), outperforming deep learning alternatives in interpretability and cost-effectiveness. By providing a deployable, low-cost diagnostic tool, this work directly supports the United Nations' Sustainable Development Goals (SDGs): SDG-3 (reducing TB mortality), SDG-9 (fostering diagnostic innovation), and SDG-10 (bridging healthcare disparities). Our results demonstrate that classical ML can rival complex models in medical diagnostics while remaining accessible to underserved populations

Keywords: *Tuberculosis,C4.5 decision tree, Sustainable health*

1. Introduction

Tuberculosis (TB) is an extremely contagious and potentially lethal condition caused by the bacterium Mycobacterium tuberculosis. It mainly affects the lungs but also has the capability to extend its reach to other organs such as the brain, spine, kidneys, and nervous system, rendering it a multidimensional and challenging health issue. Despite centuries of excellent progress in medical science TB remains one of the most significant public health issues worldwide. It is among the major causes of mortality worldwide, causing an estimated 1.5 million deaths in the year 2022 alone, according to the World Health Organization (WHO).TB has the highest burden in low- and middle-income nations, where it thrives in conditions of poverty, malnutrition, lack of good hygiene, and lack of access to proper healthcare. Furthermore, the occurrence of multidrug-resistant TB (MDR-TB) has made it difficult to control and eradicate the disease. Negative effects of tuberculosis are not only short-term; they can lead to long term sequelae such as chronic pulmonary impairment, fatigue lasting months, or even years after exposure to tuberculosis, and reduced quality of life due to successful curative treatment.

Long-term post-treatment effects have exacerbated the economic insecurity of patients and social problems like stigma and discrimination, making them less employable. The disease affects individual health but most unfavourably burdens economies and health systems, particularly in those areas where shortage is already a major challenge. A transaction is requested and initiated.

Considering the constant looming danger of TB, the situation necessitates urgent measures in evolving novel means for enhanced early detection, correct diagnosis, and improved treatment. With advancements in technology over the past few years, machine learning technology has shown great promise in tackling some of the above problems. While deep learning excels in large datasets, we included logistic regression and SVM as baselines to evaluate performance trade-offs in small-data scenarios common in TB-endemic regions. By leveraging the power of computer resources and forecasting models, researchers aim to develop more efficient and cost-

1Saintgits College of Engineering Kottayam
ORCID ID : 0009-0006-1186-6354
2 Saintgits College of Engineering Kottayam
ORCID ID : 0000-0002-9274-7608
3 Saintgits College of Engineering Kottayam
ORCID ID : 0009-0003-0339-4164
4Saintgits College of Engineering Kottayam
ORCID ID : 0009-0001-3927-8769
**Corresponding Author*
Email:helanmary.se2325@saintgits.org

effective TB diagnostic tools, especially for deployment in resource-poor settings. This study explores the potential of machine learning algorithms, such as decision trees and support vector machines, to enhance TB detection and aid in combating this chronic public health scourge globally. Through the synthesis of medical intelligence and technological progress, we intend to open up more effective responses to stem the TB burden and improve patient outcomes across the world.

2. Related Work

Tuberculosis, a global health challenge mainly in the developing world, has been the subject of research in relation to the detection of disease through the application of ML[1].It has extended beyond its old complementary role in diagnosing tuberculosis into tuberculosis triage and screening, mainly through chest radiography since the 2010s. A statistical interpretation work on the chest radiograph for the diagnosis of pulmonary tuberculosis is put forward in this study. The most significant output of this study found that detection of pulmonary tuberculosis would be done by constructing discriminant function which used maximum column sum energy texture measures where misclassification probability was less than 0.15. This study was validated, and the discriminative procedure achieved a correct classification rate of 94% [2].Logistic regression was the tool used to see the performance on tuberculosis data. According to the estimated function, pulmonary tuberculosis complications were inversely correlated with age and occupation, but favourably correlated with the patients' social history and prior exposure to TB infection. Additionally, the presence of malaria fever had an impact on the lack of pulmonary tuberculosis sequelae. Current diagnosis of TB includes smear and culture tests showing about 40% and 70% accuracy rates respectively. They have proved to be less efficient, complicated, and expensive to perform in developing countries[3].It introduces innovative computational predictive algorithms that enhance traditional decision tree methods, enabling efficient and highly accurate detection of tuberculosis (TB). These advanced techniques leverage supervised learning to classify biological samples into patient and healthy groups based on Mean Fluorescence Intensity (MFI) values of various antibodies. With accuracy levels reaching up to 94%, our approach outperforms conventional methods, offering a more reliable and effective solution for TB detection[4].In the medical field, diagnosing conditions and determining treatment plans are among the most time-intensive tasks. Specialist systems are increasingly vital for disease diagnosis due to their high accuracy in classification and diagnostic capabilities. In this study, a machine learning approach known as Support Vector Machine (SVM) was employed for the first time to assist in the initial diagnosis of tuberculosis, offering a promising tool for improving efficiency and accuracy in healthcare[5].In order to separate the intensity data over the whole area, equalization recommends showing one spreading the presumed histogram to another with wider spreading and more consistent spreading of intensity data. However, every feature in the image can have a meaning, so by using various scenarios to improve the contrast of the image, it can be seen that applying the CLAHE Algorithm's multiple layers twice can yield a very satisfactory result[6-7]. In this study, the k-nearest neighbours (KNN) algorithm will be utilized, as it is known for its strong recognition accuracy and falls under the category of supervised learning algorithms. Machine learning has gained significant traction in the medical field for analysing healthcare datasets. Research has shown that the KNN method achieved an average accuracy of 73% in detecting tuberculosis (TB) from X-ray images when combined with SURF feature extraction[8].Tuberculosis (TB) is a contagious disease that primarily targets the lungs. If not detected early, it can spread to other parts of the body and lead to severe, potentially fatal outcomes. Diagnosing pulmonary 3 tuberculosis typically involves analysing a posterior-anterior chest X-ray. The proposed method, tested on various datasets, demonstrates strong performance with an average accuracy of 95.5%, specificity of 98%, sensitivity of 93.3%, and an area under the curve (AUC) of 94.6%. These results highlight its effectiveness in identifying TB accurately and reliably[9]. Although logistic regression and SVMs are among the most popular and frequently employed algorithms for binary classification tasks they perform well with a very small dataset. On the other hand, decision trees and random forests have shown promise with a complex dataset in which relationships are usually non-linear. The k-NN algorithm has been a very successful instance-based learning algorithm. This is one of the few studies to objectively and systematically compare all these methods for their effectiveness in TB detection. It thus fills a gap that has persisted in the literature by evaluating different traditional ML algorithms on the same experimental setup and dataset. Scientists have found that logistic regression methods and SVMs are, in general, very popular algorithms for solving binary classification scenarios with robust performance shown with small data sets. Decision trees and random forests have been demonstrated to be effective with more complicated data that is non-linear in nature, and K-NN actually works pretty well in the instance-based learning problems. This is the field, however, that is unfortunately low in studies comparing any of these techniques systematically against each other in TB detection. This work hereby is intended to fill this gap in that it will compare several traditional ML

algorithms using the same dataset and experimental setup.

3. Methodology

The methodology section outlines the systematic approach adopted to conduct the research and achieve the stated objectives.

3.1. Dataset

The dataset used in this study is the Shenzhen dataset, a publicly available collection of chest X-ray images from Shenzhen No. 3 People's Hospital, China. It contains 662 images, of which 326 are labelled as TB-positive and 336 as normal. This dataset is widely recognized for its relevance to TB detection research and provides a robust basis for evaluating machine learning models[10].The positive cases contain pulmonary tuberculosis and spinal tuberculosis. Therefore, generally speaking the dataset contains 3 classes.

3.2. Preprocessing Pipeline

- Normalization: Pixel values of X-rays were scaled to [0, 1] to ensure uniformity.
- Feature Extraction: Texture features (GLCM) and intensity histograms were extracted to complement raw pixel data.
- Class Balancing: Spinal TB cases (6 samples) were augmented using Synthetic Minority Over-sampling Technique (SMOTE) to mitigate imbalance.
- C4.5 Optimization: The max depth parameter was tuned via 5-fold cross-validation (range: 1–15). Pruning was applied to minimize overfitting (confidence threshold = 0.25)

3.3. Algorithms

This section describes the details of 13 classification algorithms used for comparing performance in tuberculosis classification.

1. Decision Tree: The C4.5 decision tree algorithm has become one of the most popular techniques in providing machine learning solutions for classification problems. The working of this algorithm consists of segments of the dataset into various sub sets based on the feature that gives the maximum data gain at each step. It uses pruning techniques for the minimization of overfilling and maximizing the ability of the model to generalize unseen cases. The C4.5 algorithm is praised for its interpretability and simplicity in resource-constrained applications, such as medical diagnosis, where simple and interpretable decision rules are desired[11].
2. Logistic Regression: Logistic Regression will handle the act of classification for binary dependent variables; this itself has been popularly integrated into many non-binary classification problems as well. Logistic Regression is commended for its simplicity, interpretability, and efficiency. Hence, it finds ample acceptance in general problems such as spam detection, medical diagnosis, and customer churn[12].
3. Ada boost classifier: It is based on the boosting concept, which highlights the success of weak learners, and works incredibly well for categorization tasks. A weak learner is a model, such as a decision stump, which is a one-level decision tree that performs better than random guessing. The ensemble technique known as "boosting" trains weak classifiers one after the other by changing their weights to focus more on the incorrectly categorized samples from earlier rounds. In order to force succeeding classifiers to concentrate more on those challenging cases, AdaBoost adapts by assigning greater weight to misclassified samples in each iteration[13].
4. Gradient boosting Classifier: Gradient Boosting Classifier is an algorithm that has been used quite frequently in machine learning for classification. This is an ensemble learning method where predictions from multiple weak learners can be combined to develop a strong predictive model. The work mechanism of this algorithm relies on iteratively improving the model through the reduction of errors of previous iterations. Ensemble Learning Combines multiple models (weak learners) to produce a stronger model. Boosting is a way of training a sequence of weak models, so that each of the subsequent models corrects for the errors that the previous one has made. Gradient boosting minimizes a loss function using gradient descent by iteratively adjusting the model's parameters[14].
5. Naive Bayes Classifier: Naive Bayes is a supervised machine learning algorithm that uses probability to classify data. A probabilistic classifier based on Bayes' Theorem, assuming strong (naive) independence between features. It is highly efficient for large datasets. For continuous data, it assumes that the features are normally distributed. For discrete data, it is used in text classification[15].
6. Ridge Classifier: A machine learning technique called the Ridge Classifier was created for multi-class classification applications. By penalizing excessive weights, the Ridge Classifier, a linear classifier, prevents overfitting by adding an L2 regularization component to the loss function[16].
7. Support Vector Machine: The Support Vector Machine is one of the most powerful supervised learning algorithms that work for both regression and classification purposes, although it is relatively more widely used for classification. The primary ideology of SVM occurs when we obtain a hyperplane ideal, i.e. in a high-dimensional space separating data points coming from different classes. The optimum separation in such a case is marked by maximizing that margin, or in simple terms the distance between the hyperplane and the closest points to it (called the support vectors), thus ensuring accurate and robust classification[17].
8. Random Forest: Random Forest is a general-purpose and popular machine learning algorithm that works by learning many decision trees while training and assembling their predictions to make predictions.

Each tree is trained on a random subset of the features and data, which prevents overfitting and enhances generalization[18].

9. k- Neighbours: K-Nearest Neighbours (K-NN) is a straightforward yet powerful machine learning approach, both for classification and regression problem-solving conditions. Given an input, the K-NN method finds the 'k' nearest neighbouring data points and predicts on the basis of the majority class (for classification) or average (for regression) of these neighbours. Moreover, the algorithm is non-parametric; it does not assume any underlying distribution for the data[19].

10. Extra tree classifier: Similar to Random Forest, the Extra Trees Classifier is an ensemble learning technique that adds more unpredictability to the tree construction process. It collects the predictions from a number of decision trees[20].

Key Differences from Random Forest:

• Feature Split: Instead of finding the best split for a feature, Extra Trees chooses the split point randomly.

• Efficiency: Since it skips the costly process of finding the optimal split, Extra Trees is faster.

11. Light Gradient boosting machine: Light Gradient Boosting Machine (Light GBM) is a fast and scalable machine learning algorithm specifically built for gradient boosting jobs. It is engineered to be optimized for performance and speed, thus very ideal for processing large data sets. It further allows distributed and parallel computing, thus making it a very useful tool for classification, regression, and ranking tasks, particularly in processing big data[21].

12. Linear Discriminant analysis: Linear Discriminant Analysis, generally abbreviated LDA, is a statistical tool that can be used for classification and feature reduction. It is done by selecting the linear combination of features to class the two or more classes of data[22].

13. Quadratic discriminant analysis: The Quadratic Discriminant Analysis (QDA) build up from the Linear Discriminant Analysis (LDA). It is used to separate the data non-linearly. QDA assigns a unique covariance matrix to each class unlike LDA which offers just one covariance matrix for all classes. This makes QDA more flexible and suitable for capturing complex non-linear decision boundaries[23].

14. Extreme Gradient boosting: Boost, short for "Extreme Gradient Boosting," is a powerful machine learning algorithm built on a gradient boosting framework. It 6 relies heavily on decision trees and is widely recognized for its exceptional accuracy, scalability, and efficiency in handling large datasets[24].

4. Result Analysis

The C4.5decision tree algorithm, which is an evolution of the previous ID3 algorithm, became the model with the highest performance as regards its interpretability and balanced performance. . It is an iterative process, learning decision trees by applying a divide-and-conquer approach, partitioning datasets on the basis of the attribute which permits the highest information gain ratio. With this strategy it is possible to provide effective management of categorical, continuous, and missing data, ensuring robustness for use in a medical diagnostic scenario.

4.1. Performance Comparison

FromTable.1we identified that the decision tree is the best model that provides better performance.

Table 1. Performance Metrics of ML Algorithm

Model	Accuracy	AUC	Recall	Precision	F1
Decision Tree	0.9978	0.9979	0.9978	0.9979	0.9975
Logistic regression	0.9935	0.0000	0.9935	0.9984	0.9953
Ada boost	0.9935	0.0000	0.9935	0.9872	0.9903
Gradient boost	0.9892	0.0000	0.9892	0.9814	0.9848
K-neighbours	0.9870	0.9978	0.9870	0.9935	0.9890
Naive Bayes	0.9718	0.9723	0.9718	0.9454	0.9582
Ridge classifier	0.9718	0.0000	0.9718	0.9454	0.9582

Random Forest	0.9718	1.000	0.9718	0.9477	0.9593
Extra tree classifier	0.9718	1.000	0.9718	0.9454	0.9582
SVM	0.6639	0.0000	0.6639	0.7697	0.6720
Light Gradient	0.5965	0.9757	0.5965	0.6772	0.5382
Linear Discriminant analysis	0.4945	0.000	0.4946	0.2447	0.3274
Extreme gradient boosting	0.4772	0.9807	0.4772	0.2278	0.3064

The C4.5 decision tree achieved near-perfect classification accuracy(99.78%)with a high F1-score,demonstrating its ability to generalize well to new data. It had the lowest training time (0.147s),making it computationally efficient and ideal for resource-limited settings. Unlike deep learning methods, it provides an interpretable structure, aiding clinical decision-making. While our model shows exceptional performance, three limitations must be noted:

- Dataset Bias: The Shenzhen dataset primarily represents a Chinese population, limiting generalizability to other ethnic groups. Future work will validate the model on multi-center datasets (e.g., NIH ChestX-ray14).
- Real-World Noise: The study assumes ideal imaging conditions; accuracy may degrade with poor-quality X-rays (e.g., motion artifacts or low resolution).
- Class Imbalance: Despite SMOTE augmentation, spinal TB cases (n=6) remain underrepresented. Collaboration with TB-endemic regions is needed to collect more diverse samples.

5. Discussions

- Confusion Matrix: The dataset contains metadata (like age, sex, findings) and labels for tuberculosis (TB) detection. The labels are likely categorical, representing different diagnostic classes such as Normal, Pulmonary TB, Spinal Tuberculosis. The Figure. 1 represents the confusion matrix of proposed model. Out of 660 total images 199 were utilized as test images for the confusion matrix.

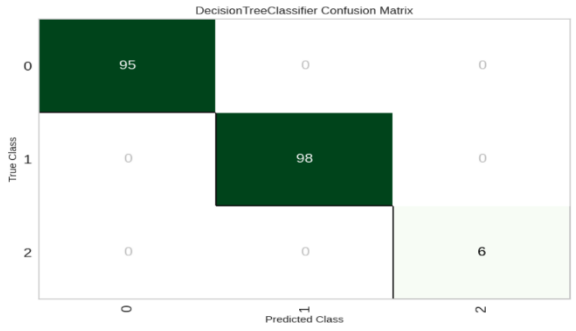

Fig 1. Confusion Matrix of Decision Tree Classifier

ROC: The decision tree classifier's ROC curve is displayed in Figure. 2 The classifier's capacity to distinguish between classes is shown by the Area Under the Curve. A perfect classifier, however, has an AUC of 1.0. Now Using AUC = 1.0 for all classes: The Decision Tree Classifier has done an excellent job of perfect classification for all the classes (class 0, class 1, and class 2). So, it can divide positive and negative samples of each class without a single error. By combining predictions, the micro-average takes into account each class's contribution equally. The performance in every class is averaged by the macro-average. Excellent overall performance is indicated by both metrics being 1.0.

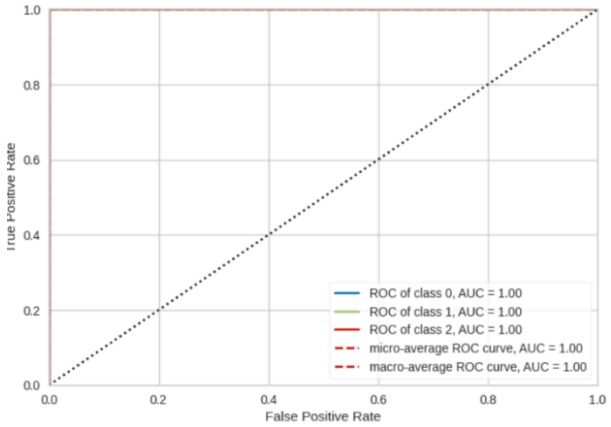

Fig 2. ROC Curve of decision tree classifier

- Dimension Analysis: t-SNE plot (perplexity=30) of feature embeddings showing clear separation between Normal (blue), Pulmonary TB (green), and Spinal TB (red) classes. Overlap between Pulmonary and Spinal TB (2% of cases) reflects shared radiographic features, suggesting future work could benefit from clinical metadata (e.g., symptom duration). The Figure. 3 shows the t-SNE Visualization.

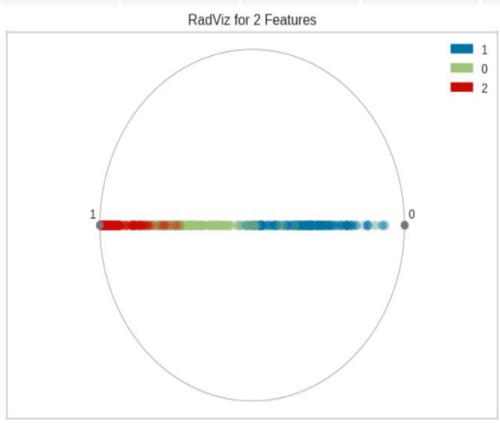

Fig. 3.t-SNE Visualization

- Learning Curve: The learning curve illustrates how the model's performance evolves as the number of training instances increases, comparing its behaviour on both training and validation data. The training score (represented by the blue line) consistently stays at 1.0, showing that the model fits the training data perfectly. On the other hand, the cross-validation score starts lower but gradually stabilizes 12 as more data is added, reflecting the model's ability to generalize to unseen data. The gap between the training and validation scores suggests a slight degree of overfitting, which is typical for decision tree models. Figure. 4 displays the learning curve for the Decision Tree Classifier, providing insights into its learning dynamics and performance trends.

Fig 4. Learning Curve of Decision Tree Classifier

- Validation Curve: Figure. 5 shows a validation curve corresponding to the Decision Tree Classifier. These curves depict the changes associated with the increase/decrease in the value of 'max depth' parameter surrounding the model performance. This explains the change in model complexity and its performance. The 'max depth' parameter tells the maximum depth of the decision tree, that is how many levels of splits can the tree make. By limiting the depth, it becomes a regularization measure and prevents the model from becoming complicated. Hence it avoids overfitting the model to the data available in training. The curve is trained by both training and cross-validation scoring, which gives an idea of how much distance the model can cover in understanding the unseen data at different levels of complexity.

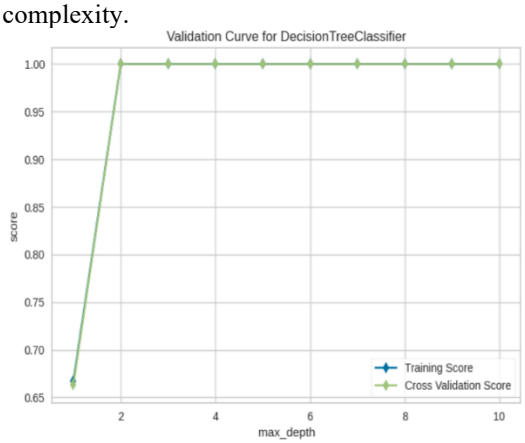

Fig 5. Validation Curve of Decision Tree Classifier

At very low 'max depth' values-i.e., one or two-the training and cross-validated scores have also been relatively low. This indicates underfitting of the model because it is a simple representation that fails to capture the patterns from the data substantially. At higher 'max depth' values, there are apparent improvements in training and validation scores. The results are close to the optimal level (close to 1.0) at

three or more depths. Notably, even high values of 'max depth' (up to 10 in this analysis) have high training and validation scores, showing that they are almost equal and implying that the model generalizes well without overfitting.

6. Conclusion

The study seriously appraised classical machine-learning techniques for the detection of tuberculosis (TB)and emphasized their merits and demerits. C4.5 Decision Tree Algorithms outperformed all other models tested in the following metrics: accuracy, precision, recall, and F1 score. The algorithm was also efficient and easily interpretable, meaning its work would be done more in tune with the hostile setting of resources in which simplicity and clarity are paramount. Random Forests and Gradient Boosting also provide competitive results but put more strain on computation and might prove impractical in resource-poor environments. Logistic regression and SVM have performed well when larger datasets were available, but these hadn't worked quite so well in smaller datasets for the present study. Logically, classical machine learning has shown promise as a cost-effective, interpretable, and accurate method for TB detection. In future work, these approaches could be integrated into real-world clinical workflows and expanded for application in field trials of larger and more diverse datasets for further improvement in robustness and generalizability.

Acknowledgements

The authors would like to express sincere gratitude to the Saintgits College of Engineering management for providing financial support for this research. Their financial assistance through research schemes:Young Research Fellowship has allowed us to get fruitful research experiences.

Author contributions

Helanmary M Sunny: Conceptualization, Methodology, Software
Dr.Anju Pratap: Reviewing and Editing
Christina Thankam Sajan: Data curation, Validation
Sandhra Merin Sabu: Visualization, Investigation,

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Hwang, E. J., Jeong, W. G., David, P. M., Arentz, M., Ruhwald, M., & Yoon, S. H. (2024). AI for detection of tuberculosis: Implications for global health. Radiology: Artificial Intelligence, 6(2), e230327.

[2] Noor, N. M., Rijal, O. M., Yunus, A., Mahayiddin, A. A., Peng, G. C., & Abu Bakar, S. A. R. (2010, November). A statistical interpretation of the chest radiograph for the detection of pulmonary tuberculosis. In 2010 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) (pp. 47-51). IEEE.

[3] Ogunsakin, R. E., & Adebayo, A. B. (2014). Performance of Logistic Regression in Tuberculosis Data. International Journal of Scientific and Research Publications, 4(9),

[4] Hussainy, S. F., Zaffar, F., Zaffar, M. A., Khaliq, A., Khan, I. H., & Ahmad, R. (2017). Decision-tree inspired classification algorithm to detect Tuberculosis (TB).

[5] Rakhmetulayeva, S. B., Duisebekova, K. S., Mamyrbekov, A. M., Kozhamzharova, D. K., Astaubayeva, G. N., & Stamkulova, K. (2018). Application of classification algorithm based on SVM for determining the effectiveness of treatment of tuberculosis. Procedia computer science, 130, 231-238.

[6] Rohmah, R. N., Handaga, B., Nurokhim, N., & Soesanti, I. (2019). A statis tical approach on pulmonary tuberculosis detection system based on X-ray image. TELKOMNIKA (Telecommunication Computing Electronics and Control), 17(3), 1474-1482.

[7] Alkhalid, F. F., Hasan, A. M., & Alhamady, A. A. (2021). Improving radio graphic image contrast using multi layers of histogram equalization technique. IAES International Journal of Artificial Intelligence, 10(1), 151.

[8] Rizal, R. A., Purba, N. O., Siregar, L. A., Sinaga, K., & Azizah, N. (2020). Analysis of Tuberculosis (TB) on X-ray image using SURF feature extraction and the K-Nearest Neighbor (KNN) classification method. Journal of Applied Information and Communication Technologies (JAICT), 5(2), 9-12.

[9] Geetha Pavani, P., Biswal, B., Sairam, M. V. S., & Bala Subrahmanyam, N. (2021). A semantic contour-based segmentation of lungs from chest x-rays for the classification of tuberculosis using Naïve Bayes classifier. International Journal of Imaging Systems and Technology, 31(4), 2189-2203.

[10] Gozes, O., & Greenspan, H. (2019, July). Deep feature learning from a hospital scale chest x-ray dataset with application to TB detection on a small-scale dataset. In 2019 41st annual international conference of the ieee engineering in medicine and biology society (embc) (pp. 4076-4079). IEEE. 15.

[11] Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 4(2), 13-19.

[12] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.

[13] Sonavane, R., & Sonar, P. (2016, December). Classification and segmentation of brain tumor using Adaboost classifier. In 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC) (pp. 396-403). IEEE.

[14] Bent´ejac, C., Cs¨org˝o, A., & Mart´ ınez-

Mu˜ noz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.
[15] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).
[16] Rakesh, K., & Suganthan, P. N. (2017). An ensemble of kernel ridge regression for multi-class classification. Procedia computer science, 108, 375-383.
[17] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.
[18] Genuer, R., Poggi, J. M., Genuer, R., & Poggi, J. M. (2020). Random forests (pp. 33-55). Springer International Publishing.
[19] Ertu˘grul, ¨ O. F., & Ta˘gluk, M. E. (2017). A novel version of k nearest neighbor: Dependent nearest neighbor. Applied Soft Computing, 55, 480-490.
[20] Shafique, R., Mehmood, A., & Choi, G. S. (2019). Cardiovascular disease prediction system using extra trees classifier.
[21] Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., & Zeng, W. (2019). Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural water management, 225, 105758. [22]
[22] Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. Nature Reviews Methods Primers, 4(1), 70.
[23] Bose, S., Pal, A., SahaRay, R., & Nayak, J. (2015). Generalized quadratic discriminant analysis. Pattern Recognition, 48(8), 2676-2684.
[24] Chen, X., Huang, L., Xie, D., & Zhao, Q. (2018). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. Cell death & disease, 9(1), 3