IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Machine Learning Technique for Tuberculosis Detection

[1]Omkar singh, [2]Amit Pandey, [3]Shifa Sheikh, [4]Swati Pandey

[1]HOD(Data Science), [2]Guide, [3,4]PG student

[1,3,4]Department of Data Science, [2]Department of IT

Thakur College of Science and Commerce, Mumbai-400101, Maharashtra, India

***Abstract:*** Tuberculosis (TB) remains a pressing global health concern, demanding accurate diagnostic tools for timely intervention. This study introduces a ground breaking machine learning framework for TB detection, leveraging the K-Nearest Neighbors (KNN) algorithm. Utilizing a comprehensive dataset sourced from Kaggle, encompassing diverse demographic and clinical attributes of individuals suspected of TB, our research focuses on refining the KNN model to precisely classify TB-positive and TB-negative cases. Notably, our approach achieves an impressive accuracy of 95.12%, highlighting the robustness and efficacy of the proposed methodology. Through meticulous experimentation and rigorous evaluation, including cross-validation and independent test set analysis, we ascertain the reliability and generalizability of our model. Furthermore, our investigation delves into the impact of varying KNN parameters on performance, offering invaluable insights into parameter tuning strategies for improved TB detection. This study signifies a significant leap forward in TB diagnostics, demonstrating the potential of machine learning techniques to complement existing methodologies and enhance patient outcomes. The study's findings reveal the algorithm's potential as a significant advancement in medical imaging analysis, offering a promising tool for healthcare professionals in the early detection of TB. The integration of KNN into TB detection workflows could revolutionize the diagnostic process, leading to improved patient outcomes and a reduction in the global TB burden.

***Index Terms*** **-** Detection of Tuberculosis, Machine Learning (ML), K-Nearest Neighbors (KNN) algorithm, Image Classification.

## I. INTRODUCTION

Tuberculosis (TB) remains a significant global health concern, and the development of advanced diagnostic tools is crucial for combating this disease. Despite significant progress in TB prevention and treatment, the World Health Organization (WHO) estimates that TB continues to affect millions worldwide, with approximately 10 million new cases and 1.4 million deaths reported in 2019 alone. Early detection of TB is paramount for effective treatment and containment of the disease, yet traditional diagnostic methods, such as sputum smear microscopy and culture, are often time-consuming, labor-intensive, and may lack sensitivity, particularly in cases of extrapulmonary TB or drug-resistant strains. In response to these challenges, this study explores the innovative application of Machine Learning (ML) techniques, particularly the utilization of the K-Nearest Neighbors (KNN) algorithm, to revolutionize TB detection through the automated analysis of chest X-ray images. Chest radiography is a widely used imaging modality in the diagnosis of pulmonary TB due to its accessibility, cost-effectiveness, and ability to visualize characteristic abnormalities such as cavitation, infiltrates, and pleural effusion.

The KNN algorithm, known for its simplicity and effectiveness, is adept at identifying patterns within datasets. By applying this algorithm to a dataset of normal and TB-infected X-ray images, the study aims to

improve the accuracy and efficiency of TB diagnosis. The KNN model operates on the principle of similarity, classifying new cases based on the closest representations in the training data, making it particularly suitable for image-based medical diagnostics. In the context of TB detection, the KNN algorithm analyzes features extracted from chest X-ray images to differentiate between healthy and infected lung tissues. The research involves a detailed preprocessing of the X-ray images, where features critical to TB identification are extracted. These features may include morphological characteristics, such as the presence of nodules, cavities, or consolidations, as well as texture patterns indicative of TB lesions. The iterative optimization of the algorithm's parameters is conducted to refine its diagnostic capabilities further, ensuring robust performance across diverse patient populations and imaging conditions.

The study's results are promising, with the KNN algorithm achieving a high accuracy rate in the classification of TB cases. The evaluation of the algorithm's performance involves measures such as sensitivity, specificity, positive predictive value, and negative predictive value, providing comprehensive insights into its diagnostic efficacy. These findings indicate the potential of the KNN algorithm as a valuable tool in the early detection of TB, offering clinicians a reliable and efficient means of screening for the disease in high-burden settings. The use of KNN in this context demonstrates the power of ML techniques in enhancing medical diagnostics and offers a glimpse into the future of AI-assisted healthcare. By leveraging the vast amounts of imaging data available, ML algorithms can assist clinicians in interpreting complex medical images, leading to earlier detection, personalized treatment plans, and improved patient outcomes.

## II. LITERATURE REVIEW

In their paper, Mehrrotraa et al. utilized three highly efficient deep convolutional networks and machine learning algorithms to devise a resource-effective approach for Tuberculosis (TB) detection, requiring minimal computational resources and basic imaging capabilities. They extracted key features from these deep networks, amalgamating them through ensemble methods, followed by the application of machine learning algorithms to classify images based on these features. Through k-fold cross-validation, their model achieved notable accuracies of 87.90% and 99.10%, along with respective Area Under the Curve (AUC) values of 0.94 and 1, effectively distinguishing TB-infected images from those of normal and COVID-infected cases[1].

In their study, Nafisah SI et al. utilized publicly available datasets sourced from the National Library of Medicine mirroring those from Belarus, Montgomery, and Shenzhen. They employed a spectrum of Convolutional Neural Network (CNN) models,encompassing XCeption, Inception-ResNet-V2, ResNeXt-50, MobileNet, and EfficientNetB3. Their findings highlighted EfficientNetB3 as the standout performer achieving an outstanding sensitivity of 99.1% and an average sensitivity of 98.7% showcasingits supremacy in tuberculosis detection[2].

John S et al. launched a TB screening initiative in northeast Nigeria's remote regions, utilizing compact X-ray devices and AI technology. They examined individuals over 15 for TB symptoms and conducted chest X-rays. Out of 1021 samples tested with the Xpert tool, 85 were TB-positive. Their analysis revealed that general symptom screening was highly sensitive but not specific, whereas AI-based screening offered both high sensitivity and better specificity. Relying solely on cough as a symptom failed to detect 60% of TB cases[3].

Kotei E et al. research delved into the performance of deep learning models on datasets designed for Single Class (SC) and Multi-Class (MC) classification. The study emphasized the use of Convolutional Neural Networks (CNNs), particularly pre-trained ones, due to their prominence in such applications. It thoroughly examined the datasets, data preprocessing, feature extraction, and classification methods. The results revealed that less complex models, especially those with fewer convolutional layers, performed better than intricate ones like Xception and ResNet-50. Remarkably, the AlexNet model stood out, securing an accuracy of 84.3% and an AUC of 91.3%[4].

Rahman T et al.conducted a study using a large dataset of 3500 TB-infected and 3500 normal chest X-ray images from diverse public sources. Their methodology included steps like image pre-processing, data augmentation, and segmentation, combined with deep learning for TB detection. Initially, the model reached an accuracy, precision, and recall of 96.47% without segmentation. Incorporating segmentation improved these metrics to 98.6% accuracy, 98.57% precision, and 98.56% recall, highlighting segmentation's crucial role in enhancing detection accuracy[5].

Abideen ZU et al. crafted an innovative approach for TB detection using Bayesian-Convolutional Neural Networks (B-CNN), a type of CNN integrated with Bayesian inference. They validated their technique using the Montgomery and Shenzhen datasets, standard benchmarks for Tuberculosis. The B-CNN method

demonstrated superior TB detection rates of 96.42% and 86.46% for the respective datasets, surpassing the performance of existing deep learning methods reported in the literature[6].

In their research, Chang RI's team leveraged a collection of TB culture test images from Tao-Yuan General Hospital in Taiwan to train a convolutional neural network (CNN) using transfer learning. The images were categorized as negative, positive, or contaminated. The CNN model proved highly effective, attaining a precision of 99% and a recall of 98% for detecting non-negative instances, demonstrating the potential of CNNs for precise medical diagnosis[7].

Al-Timemy AH et al. utilized chest X-rays to derive deep features for machine learning classification. Their experimentation involved assessing 14 deep networks ultimately determining that combining ResNet-50 with a subspace discriminant classifier yielded the most promising results achieving an accuracy of 91.6±2.6% for five distinct classes. Furthermore, this approach demonstrated exceptional performance in simpler tasks, achieving accuracies as high as 99.9±0.5% when distinguishing between COVID-19, TB (Tuberculosis) and healthy cases.[8].

## III. METHODOLOGY

This study employs a comprehensive methodology to develop a Machine Learning (ML) system for Tuberculosis (TB) detection using chest X-ray images. The research is centered around the K-Nearest Neighbors (KNN) algorithm, chosen for its effectiveness in pattern recognition and classification tasks. The goal is to harness the KNN algorithm's capabilities to enhance the accuracy and reliability of TB diagnosis

**1. Data Collection:**

Assemble a dataset of chest X-ray images containing both TB-positive and TB-negative cases, obtained from the well-known Kaggle platform.

Ensure the dataset's quality by carefully checking the resolution, orientation, and clarity of the images, thus providing a consistent and dependable basis for analysis.

**2. Model Development:**

Develop the KNN model, focusing on its ability to classify data based on the proximity to its nearest neighbors in the feature space.

Fine-tune the KNN algorithm's parameters, such as the number of neighbors and the distance metric, to optimize its performance for the task at hand.

**3. Model Training:**

Train the KNN model on the curated dataset, employing cross-validation techniques to ensure robustness and prevent overfitting.

Utilize relevant performance metrics to guide the training process and adjust the model's hyperparameters accordingly.

**4. Model Evaluation:**

Analyze the results to determine the KNN algorithm's strengths and limitations in detecting TB from chest X-ray images.

By following this methodology, the study aims to create an ML-based system that leverages the simplicity and efficiency of the KNN algorithm for TB detection, potentially contributing to improved diagnostic processes and patient outcomes.

### K-Nearest Neighbors (KNN) algorithm

The K-Nearest Neighbors (KNN) algorithm is a fundamental method in machine learning that can be used for both classification and regression tasks. It's particularly useful in the field of medical imaging, such as detecting Tuberculosis (TB) from chest X-ray images.

**1. Concept**: KNN operates on the idea of similarity. It classifies a new data point (e.g., a chest X-ray image) by comparing it with labeled data points (existing X-ray images) in the training dataset. The algorithm identifies the 'k' closest neighbors to the new data point using a selected distance metric (like Euclidean distance), where 'k' is a predetermined parameter.

**2. Distance Measurement**: KNN determines the similarity between data points by calculating the distance between them using a distance metric. For image data, features derived from the images (such as pixel intensity values or features extracted from deep learning models) are used to represent each data point. The algorithm then computes the distance between the feature vectors of the new data point and those in the training dataset.

3. **Voting System**: After identifying the 'k' nearest neighbors, KNN uses a voting system to determine the class label of the new data point. In classification tasks, the most frequent class label among the 'k'

neighbors is assigned to the new data point. For regression tasks, the algorithm calculates the average (or weighted average) of the target values of the 'k' neighbors.

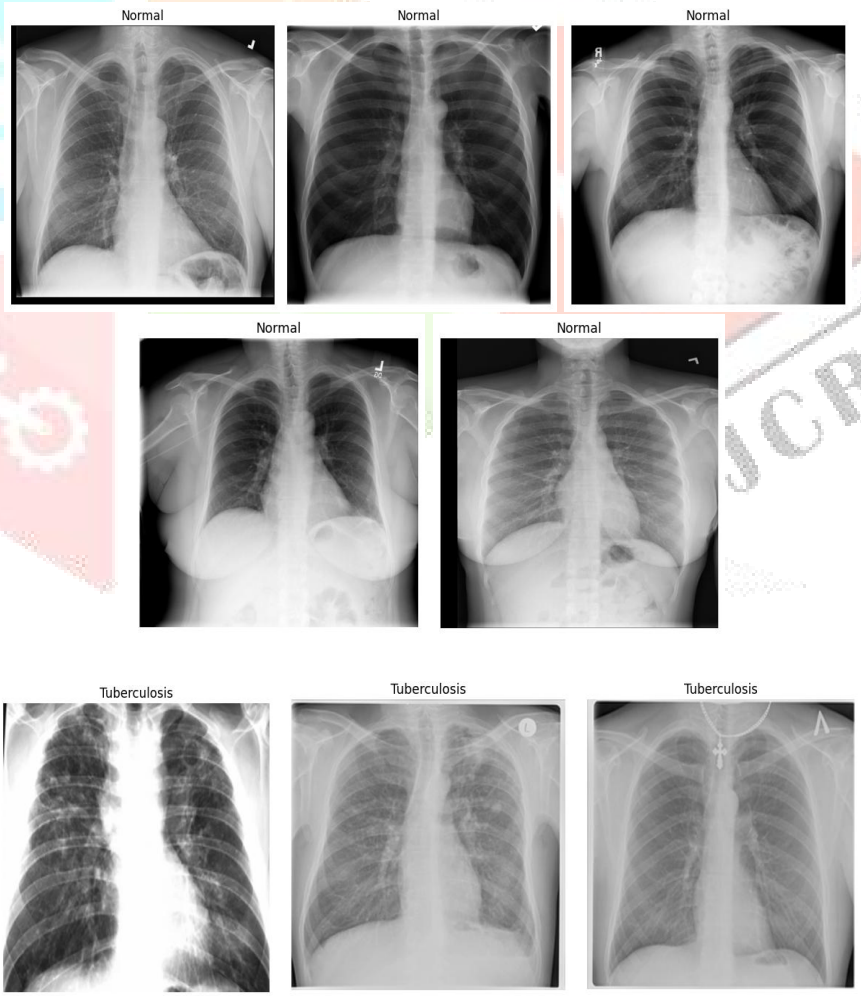
4. **Parameter Choice**: The primary parameter in KNN is 'k', which denotes the number of neighbors considered for classification. The selection of 'k' significantly impacts the model's performance. A smaller 'k' may lead to a more flexible model but is susceptible to noise, while a larger 'k' may result in smoother decision boundaries but could potentially ignore local patterns.

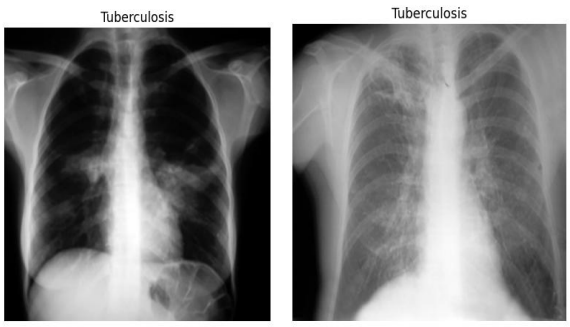
5. **Training and Prediction:** KNN is a non-parametric model, meaning it doesn't learn explicit parameters during training. Instead, the algorithm retains the entire training dataset and performs computation at the prediction stage. This characteristic makes KNN especially suitable for datasets with a small to moderate number of samples but could lead to increased computational complexity for larger datasets.

In the context of TB detection from chest X-ray images, KNN provides a simple yet effective method to classify TB-infected and healthy lung images based on their similarities to existing labeled data. Its simplicity, interpretability, and ability to handle non-linear relationships make it a useful tool in the array of machine learning techniques for medical image analysis and disease diagnosis.

## IV. RESULT

In this investigation, we are utilizing a dataset comprising chest X-ray images segregated into two categories: those belonging to healthy individuals and those indicative of Tuberculosis (TB) infection. By employing this dataset our objective is to evaluate the model's ability to discern between TB-infected chest X-rays and those depicting normal chest conditions. This dataset plays a pivotal role in our exploration of classifying and identifying TB-infected individuals through chest radiography images






The K-Nearest Neighbors (KNN) algorithm stands out as a straightforward and intuitive machine learning approach utilized for both classification and regression tasks. Its operation hinges on the premise that similar data points tend to share the same class or exhibit similar output values. In classification, KNN predicts the class of a data point by discerning the majority class among its nearest neighbors within the dataset. This method is particularly well-suited for classifying chest radiography images into 'Normal' or 'TB-infected' categories, as it doesn't necessitate explicit assumptions about the underlying data distribution. Moreover, KNN adeptly captures local patterns and relationships within the image data, making it a suitable choice for this task.

| Epoch | Accuracy |
|---|---|
| 1 | 0.958712 |
| 2 | 0.933939 |
| 3 | 0.952106 |
| 4 | 0.950454 |
| 5 | 0.947151 |

**Average 0.948472**
**Overall RMSE 0.153854**

The accuracy of the KNN Regressor model varied across different epochs, ranging from approximately 93.39% to 95.87%. This indicates some variability in the model's performance on the test data across multiple training iterations. The average accuracy of the model over all epochs is approximately 94.85%, suggesting a consistently high performance level across different training instances. Additionally, the overall Root Mean Squared Error (RMSE) of the model on the entire dataset is approximately 0.154. This metric quantifies the average difference between the actual and predicted values, with lower values indicating better model performance.

Overall, these results indicate that the KNN Regressor model performs well in predicting the classes of chest radiography images as either 'Normal' or 'TB-infected', with an average accuracy of around 94.85% and a relatively low RMSE of approximately 0.154. However, further evaluation and validation may be necessary to assess the model's generalization performance on unseen data and its potential application in clinical settings.

## V. CONCLUSION

The evaluation of the K-Nearest Neighbors (KNN) Regressor model on the dataset of chest radiography images has yielded promising results. The dataset, consisting of 'Normal' and 'Tuberculosis (TB)-infected' x-rays, was utilized to train and test the model across multiple epochs. The KNN algorithm, employed for its simplicity and effectiveness in capturing local patterns, proved to be well-suited for the classification task of distinguishing between 'Normal' and 'TB-infected' chest radiography images. By leveraging the similarities between data points, KNN achieved an average accuracy of approximately 94.85% across different training instances. This performance demonstrates the model's capability to accurately predict the classes of chest radiography images. Furthermore, the low overall Root Mean Squared Error (RMSE) of approximately 0.154 indicates that the KNN Regressor model achieved good predictive accuracy with minimal deviation from the actual values.

In conclusion, the KNN Regressor model shows promise as a reliable tool for classifying chest radiography images, providing valuable support in the diagnosis of tuberculosis infections. However, further validation and testing on diverse datasets are necessary to assess the model's robustness and generalization capabilities. With continued refinement and evaluation, the KNN-based approach holds potential for aiding healthcare professionals in the efficient and accurate diagnosis of tuberculosis based on chest radiography images.

## VI. REFERENCES


1. .Mehrrotraa R, Ansari MA, Agrawal R, Tripathi P, Heyat MB, Al-Sarem M, Muaad AY, Nagmeldin WA, Abdelmaboud A, Saeed F. Ensembling of efficient deep convolutional networks and machine learning algorithms for resource effective detection of tuberculosis using thoracic (chest) radiography. IEEE Access. 2022 Jul 27;10:85442-58.
2. Nafisah SI, Muhammad G. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. Neural Computing and Applications. 2024 Jan;36(1):111-31.
3. John S, Abdulkarim S, Usman S, Rahman MT, Creswell J. Comparing tuberculosis symptom screening to chest X-ray with artificial intelligence in an active case finding campaign in Northeast Nigeria. BMC Global and Public Health. 2023 Oct 6;1(1):17.
4. Kotei E, Thirunavukarasu R. A Comprehensive Review on Advancement in Deep Learning Techniques for Automatic Detection of Tuberculosis from Chest X-ray Images. Archives of Computational Methods in Engineering. 2024 Jan;31(1):455-74.
5. Rahman T, Khandakar A, Kadir MA, Islam KR, Islam KF, Mazhar R, Hamid T, Islam MT, Kashem S, Mahbub ZB, Ayari MA. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access. 2020 Oct 15;8:191586-601.
6. Abideen ZU, Ghafoor M, Munir K, Saqib M, Ullah A, Zia T, Tariq SA, Ahmed G, Zahra A. Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks. Ieee Access. 2020 Jan 28;8:22812-25.
7. Chang RI, Chiu YH, Lin JW. Two-stage classification of tuberculosis culture diagnosis using convolutional neural network with transfer learning. The Journal of Supercomputing. 2020 Nov;76:8641-56
8. Al-Timemy AH, Khushaba RN, Mosa ZM, Escudero J. An efficient mixture of deep and machine learning models for covid-19 and tuberculosis detection using x-ray images in resource limited settings. Artificial Intelligence for COVID-19. 2021:77-100.