CHAPTER 1
INTRODUCTION

1.1 Background

Tuberculosis (TB), a prevalent acute respiratory infection impacting the alveoli and distant air passages, poses a significant global health challenge, linked to elevated morbidity and both short- and long-term mortality across all age demographics globally [1], [2], [3]. Tuberculosis can be caused by a wide range of microorganisms, such as bacteria, respiratory viruses, and fungi. The prevalence of these microorganisms varies significantly across different geographic regions [4], [5], [6]. Tuberculosis has another type, but on this study focused in pulmonary tuberculosis that infected just infected through lungs. Prompt identification of Tuberculosis followed by the administration of suitable medication can [7], [8]. The latest Tuberculosis detection is highly needed, as it becomes faster and more accurate to detect the disease. Therefore, it will expedite further examination and decision-making processes by medical professionals using Computer-Aided Diagnosis (CAD) [9], [10], [11]. Computer-Aided Diagnosis (CAD) is a technology that employs artificial intelligence and image processing techniques to assist doctors in diagnosing diseases [12], [13], [14], [15]. AI plays a critical role in CAD, utilizing machine learning (ML) algorithms to develop models capable of recognizing anomalies or relevant features from medical data.

In the realm of machine learning, various techniques are employed to strengthen diagnostic capabilities, one of which is ensemble learning, like the Voting Classifier. Voting classifier machine learning is an ensemble technique that employs voting rules over a set of randomly-generated classifiers to discover the optimal classifier for a dataset without requiring deep domain expertise [16], [17], [18] . Voting classifiers in supervised learning leverage voting mechanisms among instances or multiple voters to categorize new observations, emphasizing the importance of error tolerance for mission-critical applications [19], [20], [21]. In bioinformatics research, the integration of ensemble and deep learning methodologies into ensemble deep learning enhances the precision, robustness, and

replicability of models [22]. Ensemble learning algorithm improves classification performance in imbalanced data. Du *et al*.[23].

Rahman *et al.* [24] employed nine different deep convolutional neural networks (CNNs), including ResNet18, ResNet50, ResNet101, ChexNet, InceptionV3, Vgg19, DenseNet201, SqueezeNet, and MobileNet, for transfer learning with pre-trained initial weights. These CNNs were trained, validated, and tested to classify TB and normal cases. The study conducted three experiments: X-ray image segmentation using two U-net models, classification using X-ray images, and classification using segmented lung images. However, classification using segmented lung images surpassed that of whole X-ray images, with DenseNet201 achieving an accuracy, precision, sensitivity, F1-score, and specificity of 98.6%, 98.57%, 98.56%, 98.56%, and 98.54%, respectively. The study also utilized visualization techniques to confirm that the CNN predominantly learns from segmented lung regions, resulting in higher detection accuracy. The proposed method, exhibiting state-of-the-art performance, holds promise for computer-aided faster diagnosis of tuberculosis.

Natarajan *et al.* [13] create a system called tbXpert was proposed, utilizing deep learning methods, specifically Deep Fused Linear Triangulation (FLT) to reconcile these variations and similarities. The system processes CXR images without requiring segmentation and trains on deep fused images using a deep learning network with residual connections. Trained on a large dataset of 3500 TB and 3500 normal CXR images, the model achieved impressive results with an accuracy of 99.2%, sensitivity of 98.9%, specificity of 99.6%, precision of 99.6%, and an AUC of 99.4%. This suggests that tbXpert can be a valuable tool for computer-aided diagnosis, reducing the time and effort required by radiologists and minimizing reliance on the expertise level of specialists.

Alqatahni *et al.* [25] explores the efficacy of utilizing a pre-trained network in conjunction with oversampling techniques for tuberculosis (TB) classification, and contrasts the findings with recent research employing the same dataset. The dataset comprises 3500 uninfected TB cases and 700 TB-infected cases. To address class imbalance, the study employs oversampling techniques on X-ray TB images

before feeding them into multiple pre-trained networks for TB classification. The oversampling technique significantly improves TB classification performance compared to other reported pre-trained models. Notably, Inceptionv3 demonstrates promising results, achieving 99.94% accuracy, 99.88% precision, 100% recall, and 99.94% F1-Score.

Geethamani *et al.* [26] developed machine learning and deep learning models are explored for diagnosing diseases early to prevent long-term effects. The Random Forests classifier achieved an accuracy, precision, recall, and F1 score of 97% each. Augmentation techniques were utilized to improve accuracy, and Histogram of Oriented Gradients (HOG) was employed as a feature extraction method. As a result, the proposed model RF-HOGADM is deemed most suitable for TB detection from chest X-ray images.

Jauhari *et al* [9] developed a severe tuberculosis (TB) detection system using the TB dataset with 4200 data points (3500 Normal and 700 TB). It aimed to create a lightweight computation system using a Voting Classifier in Ensemble Learning as the classifier for imbalanced data. Initial experiments used the SVM and Random Forest models separately, achieving accuracies of 98.6% and 98%, respectively. These models were then combined using Ensemble Learning without feature extraction. The results showed that the accuracy, AUC, Recall, Precision, and F1-score of the Voting Classifier reached 99.1%, 99.3%, 99%, 98%, and 98%, demonstrating a significant improvement in performance.

Table 1 Previous study

Study	Method	Database	Performance
Rahman *et al.*[24]	Image Segmentation, & Deep Learning (9 pre-trained CNN).	NLM (MC &CHN), Belarus, NIAD TB Potal & RSNA	Accuracy 98.6%
Natarajan *et al.* [13]	MobileNet and Transfer Learning	NLM & NIH	Accuracy 99.2%
Alqatahni *et al.* [25]	Inceptionv3 with Transfer Learning (Imbalanced dataset)	NLM (MC &CHN), Belarus, NIAD TB Potal & RSNA	Accuracy 99.94%
Geethamani *et al.* [26]	HOG with RF-HOGADM	NLM (MC &CHN), Belarus, NIAD TB Potal & RSNA	Accuracy 97%
Jauhari *et al* [9]	Plain dataset with Voting Classifier	NLM (MC &CHN), Belarus, NIAD TB Potal & RSNA	Accuracy 99.16%

Consequently, to facilitate this study, ensemble learning is employed to address the data imbalance issue without the necessity of accessing datasets requiring permissions. Imbalanced datasets frequently appear in practical classification problems, such as identifying oil spills in satellite radar imagery or diagnosing medical conditions [27]. Also, ensemble learning with voting classifier (machine learning) is more efficient for all computer. According to Taye in *Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions*,Taye n.d *et al* "Machine learning programs are typically less complicated than deep learning algorithms and may frequently be executed on standard computers" [28]. This highlights the accessibility and lower

computational requirements of machine learning models compared to deep learning, which often necessitates specialized hardware for optimal performance. The issue of data imbalance occurs when one class in a dataset has significantly more samples than another. In the medical context, this often happens when there are many more healthy patient records compared to those with a particular disease. This imbalance can lead to bias in machine learning models, where the model becomes more accurate at predicting the majority class (healthy patients) but less effective at identifying the minority class (patients with the disease). As a result, even if the model shows high overall accuracy, its performance in detecting disease cases can be very poor. This is particularly risky in medical applications where failing to detect a disease can have serious consequences. To address this issue, techniques such as oversampling the minority class, undersampling the majority class, and using appropriate evaluation metrics should be applied to ensure the model can accurately and fairly recognize and predict both classes.

1.2 Problem Identification

The main issue identified is the lack of exploration and utilization of Ensemble Learning techniques, particularly the Voting Classifier, for reliable tuberculosis (TB) detection using chest X-ray images. Although Deep Learning methods have shown promising results, they may not always be practical due to data limitations and resource constraints. Additionally, the computational demands of Deep Learning are heavier compared to Machine Learning methods when using certain parameters. Computational efficiency is crucial in the biomedical field with lighter computation, the tools used for image detection become faster and more effective. With imbalanced datasets, it will be easier to update the system for a wider variety of variations.

Machine learning is characterized by its light computational requirements and ability to achieve high accuracy, making it suitable for faster processing in resource-constrained environments. On the other hand, deep learning offers exceptional accuracy but demands significant computational resources, which may be a limitation in certain contexts. By choosing machine learning, the classification task

focuses on balancing computational efficiency with accuracy, providing a practical and reliable solution for TB detection. (Shown in Figure 1)

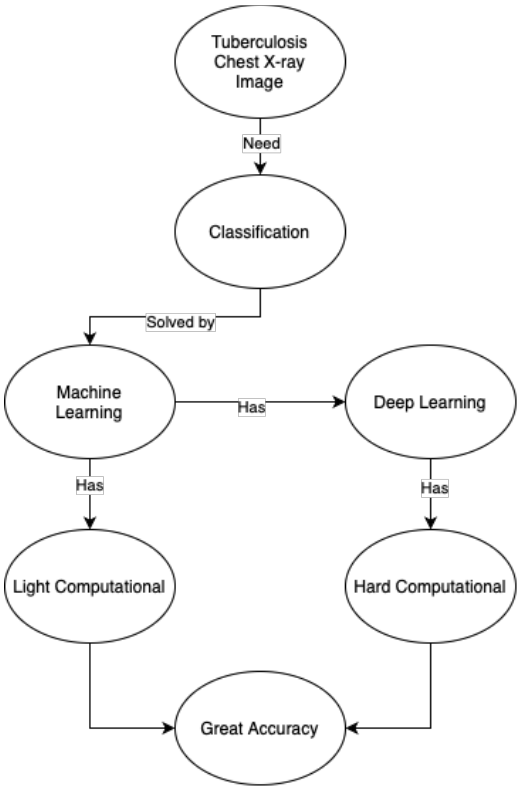

Figure 1 Problem Identification Mindmap.

1.3 Objectives

The Based on the issues identified in the research conducted, the following are the objectives of this study:

1. Utilizing a dataset accessible from previous research, consisting of 3500 samples (Non-Tuberculosis) and 700 samples (Tuberculosis), which constitute an imbalanced dataset.
2. Comparing Single Machine Learning with Ensemble Learning (Voting Classifier) as the classifier and previous study.
3. Creating lightweight computation with good system performance, making the tools used for image detection more efficient.
4. Achieving classification results more than non-imbalance classification previous study.

1.4 Scope of Work

To ensure this research does not deviate from the requirements stated, the scope of work are as follows:

1. Created an updated algorithm using python programming language-based system for solving Tuberculosis detection using imbalance dataset;
2. Implementing the proposed algorithm;
3. Conduct simulations and compare the proposed algorithm's performance in Tuberculosis classification;
4. Implementing the proposed algorithm with the improvement
5. Hardware will not be discussed.