Enhancement of Tuberculosis Detection Using Ensemble Classifier with Quadtree Method: A Preliminary Study

Chong Joon Hou[1], Laura P. Jack[1,*], Aslina Baharum[1] and Noorsidi Aizuddin Mat Noor[2]

[1]*UXRL, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia*

[2]*UTM CRES, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*

Abstract: Tuberculosis is an infectious disease caused by a bacillus called Mycobacterium tuberculosis. It can lead to death in untreated and inappropriately treated patients. An early diagnosis of the disease not only improves treatment success but also reduces death rates. Lung region is the most affected part of Tuberculosis and the process of medical image classification is still carried out manually using the knowledge of the physician or radiologist, which leads to inaccurate and slow process of TB identification. Therefore, this study proposed to enhance tuberculosis detection using a different combination of machine learning and image processing methods on the image dataset.

1. INTRODUCTION

The classification and detection of tuberculosis (TB) is a challenge for doctors and radiologists in the medical field. The faster detection of TB will be able to increase the chance for the treatment of TB since the earlier detection of the TB diseases, the earlier the treatment can be done and the higher the survivability. Many approaches can be used to identify or detect the TB. As the traditional technique, tuberculin skin test was used to screen for latent TB infection and later Chest X-Ray (CXR) has been introduced to find abnormalities or TB of a patient faster and efficiently. Because of artificial intelligence technology, CXR images have been used to detect TB. Image classification methods have been used widely recently to process and analyze the images in detail through computational methods and are able to produce satisfying results from prediction and classification.

The traditional method was time consuming in obtaining the results, either positive or negative TB. However, current image classification methods focus on extracting essential features and recognize complex patterns from images and use it as the key to classify or predict the input of images belonging to which category. Existing methods Support Vector Machine (SVM) [1-4], K-Nearest Neighbor (K-NN) [2, 3], and Multilayer Perceptron (MLP) are machine learning algorithms that are widely used for solving pattern recognition problems. Convolutional Neural Network (CNN) is another approach to analyze images and

capture the important information of images and use it to classify correctly input images into respective categories. CNN extracts features automatically while the other methods mentioned earlier extract manually by defining features to extract. The performance of CNN has been improved as having deeper layers. However, CNN with deep structure can cause overfitting problems. Not only have all these methods been used on detecting and classifying the existence of Tuberculosis from CXR images, the methods have yielded satisfying accuracy with more than 90%.

In order to improve classification or prediction accuracy, a method of combining different models called ensemble method is proposed. With this proposed method, the accuracy of prediction will be based on overall mean accuracy of different models (SVM, K-NN and MLP) and (CNNs). If either one model gives bad accuracy, this proposed method can average the overall mean found from each individual model and eventually the accuracy from each model is unified and voted at last [5]. In addition, an image representation method called Quadtree is also proposed to improve the segmentation process of images to reduce the number of grid (Pixel) that were too many in a regular grid image. A quadtree is based on the successive subdivision of a region into four equally sized quadrants and paths generated by the normal grid are suboptimal. Quadtrees allow efficient partitioning of the environment since single cells can be used to encode large empty regions [6]. The proposed method in this research will generate a path on the important features from CXR images instead of generating every single pixel on the image. Section 2 introduced some related works concepts to the machine learning model, CNN model, and quadtrees method. The method proposed

in this paper is explained in section 3 and the results & discussion are discussed in section 4. The last section is about the conclusion and recommendation of this proposed research.

2. RELATED WORKS

Many studies and researches have been done in the image classification field. One of the most important image classification in the medical field is classification of CXR into two different categories which are normal and TB detection where the traditional method is based on laboratory experiment results which is not suitable for the method of detection nowadays. Advancement in technology has introduced artificial intelligence as one of the problem solving methods. This method has been used widely and efficiently in recent years in many different fields such as computer science, engineering, astronomy, medical and etc. [7] have applied image classification by developing an image analysis system which functions to determine the type and quantity of foods and beverages based on images that are acquired by using a mobile device. This system is able to categorize the foods and beverages through detection of food and beverage items in the image. Another classification research on images has been done by (Treebupachatsakul & Poomrittigul, 2019) [8]. They have conducted research on bacteria classification using image processing and deep learning approach - Artificial Neural Network (ANN) and CNN and aimed to classify bacteria species with shorter analyzing time and increase accuracy of diagnosis process. They have undergone experiment on high resolution dataset images and standard resolution dataset images respectively, and the results concluded that recognition accuracy is higher when using high resolution images as dataset. This showed that quality of images is one of the reasons to have better classification models. They finally achieved accuracy of 96% however their limitation is limited to only two species of bacteria.

Next, (Mathews & V.R, 2015) [1] has conducted research on detection of tuberculosis using CXR with SIFT algorithm as segmentation method together with SVM as classification model. They used SIFT descriptor which is considered the best performing local feature descriptor to model the local gradient information of the CXR images. The results showed that the SIFT algorithm provides an efficient lung boundary detection of images and this method improves the overall feature extractions process and gives a high accuracy of 95% in determining TB in

CXR. Another study done by (Antony & K, 2017) [2] also proved that using artificial intelligence algorithms - (SVM), (K-NN) classifier on detection of TB using CXR obtained high accuracy. They applied methods like Gaussian filter, median filter on the X-Ray images to remove unwanted noise and aid to get fine textural features. This indicates that enhancing images will help in the feature extraction process. Finally their proposed method gave an over 80% accuracy in identifying the positive TB images. CNN has become more popular in the object and image classification field as this method is widely used as a research method on these fields. (Paras Lakhani, 2017) [9] focused on studying automated classification of pulmonary tuberculosis by using CNN. They used CNN architecture AlexNet and GoogleNext on CXR and obtained overall 86% accuracy. (Liu, *et al*., 2017) [10] proposed a novel method using CNN to deal with unbalanced, less-category X-ray images. They have studied and evaluated various CNN architectures and training parameters for TB X-ray image dataset, then applied the suitable transfer learning techniques in chest X-ray images. After that, finetuned the pre-trained CNN model (ImageNet) from natural image dataset to their medical X-ray image dataset to detect TB and find its effectiveness and efficiency than training from scratch. The final classification accuracy of CNN models yield 85.68%. Another approach called ensemble method has been applied by (Melendez, *et al*., 2015) [3] on combining K-NN and SVM in detection of tuberculosis on CXR. Their experiment managed to achieve accuracy of 86% from a total of 917 samples images. Next, (Rusdah, Winarko, & Wardoyo, 2015) [4] used two different dataset RM2_17 datasets and RM2_14_2C dataset respectively on 4 single classifier (C4.5, Naive Bayes, Backpropagation, and SVM) and ensemble classifier (Bagging and Boosting). These two datasets are trained and tested on each single classifier and as well as each ensemble classifier and accuracy obtained from each classifier is recorded. The accuracy obtained shows that the ensemble classifier produced better performance in terms of accuracy than using a single classifier.

Image representation play and important role in segmentation and feature extraction process as [11] concluded that the Quadtree Segmentation algorithm produced segments that were annotated 50% better than those produced with grid segmentation and 35% better than those generated with normalized cuts. In addition, the proposal segments are about 17 times faster than normalized cuts which are mentioned by

(Marquez, Escalante, & Sucar, 2010) [11]. They also mentioned that proper segmented images will improve the feature extraction process as the quadtree method will also enhance the image by drawing quadrants on lines or important features on images. This allows important information on images being focused and useful less information being neglected. (Deivanayagampillai, 2017) [12] mentioned quadtree decomposition able to define regions of interest for further processing as image segmentation, feature selection, object detection, sample annotation and etc. He also mentioned that smaller blocks describe fine details more accurately and bigger blocks incorporate similar regions to represent them with fewer information as possible.

3. METHODOLOGY

3.1. Phase 1

In this phase, data preparation, image pre-processing and quadtree segmentation will be perform. In data preparation stage, samples of chest x-ray is acquired and then undergo pre-processing to enhance the quality of images. The end of this phase is by conducting segmentation using proposed quadtree method to separate each image into smaller pieces for phase 2 process.

3.2. Phase 2

In this phase, segmented images is then undergo feature extraction process here. These extracted features is then used in training process for 3 different classifier.

3.3. Phase 3

Final phase is the prediction and evaluation stage where each built model is use to predict the input image. The accuracy of each model will be calculated and evaluated.

4. RESULT AND DISCUSSION

4.1. Data Acquisition

The Chest X-Ray images are obtained from two different online databases called ShenZhen dataset and Montgomery dataset. ShenZhen dataset was created by U.S. National Library of Medicine (USNLM) in association with Guangdong Medical College, ShenZhen, China. It consists of 662 CXRs with 336 TB manifested CXRs. Whereas Montgomery dataset was created by (USNLM) using the services of the health department at Montgomery County (MC), USA. It consists of 138 with 58 CXRs consisting TB. The size of all of ShenZhen images are 3000*3000 pixels
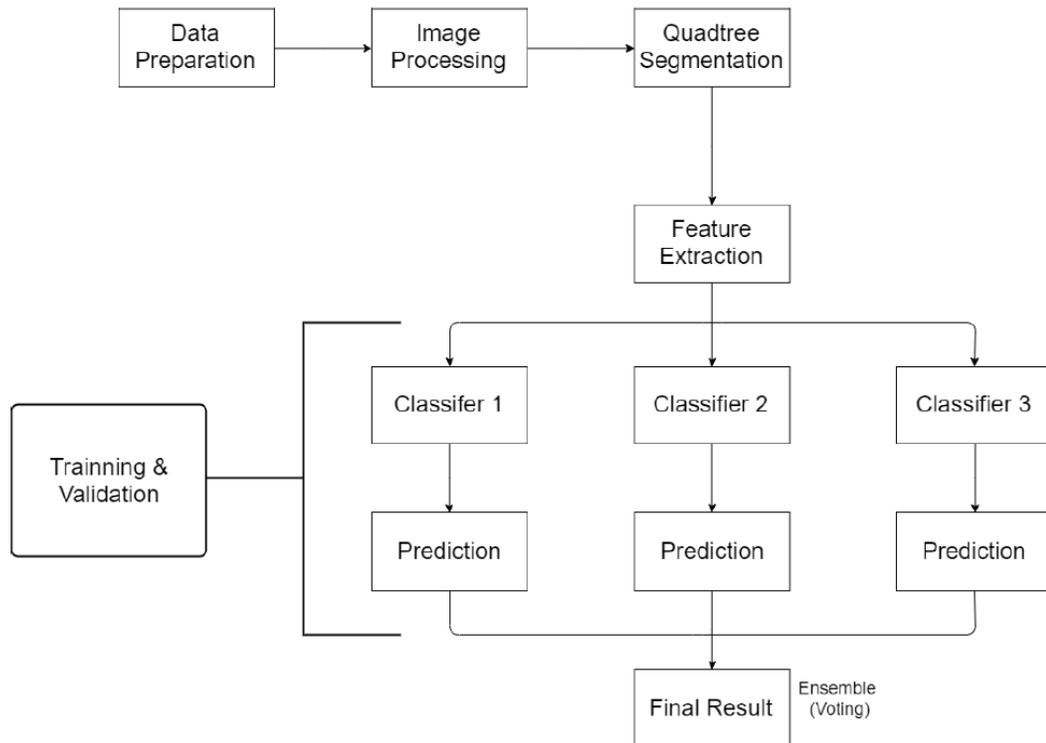

Figure 1: Flow of methodology.

whereas image size for Montgomery dataset are 4020*4892 or 4892*4020 pixels. These X-ray were captured using Eureka stationary X-Ray Machine (CR) and are presented in Portable Network Graphic (PNG) format as 12-bit gray level images.

Table 1:   Quantity of Images in each Dataset

Category	ShenZhen	Montgomery	Total
Non-TB	80	326	406
TB	58	336	394

4.2. Image Processing

Image processing stage is necessary to enhance the quality of images to enable it to undergo extraction to extract meaningful information from it by performing some operation on the images. Image processing methods treat the image as a two-dimensional matrix with discrete value for particular position as well as intensity. Pixels indicate the position and the value of the pixel represents the intensity at that position. It involves a series of steps that necessary to be done to extract the desired information from the digital images (CXR). Various pattern recognition problems, the techniques involved may differ. In the step of image processing for this research, the CXR images obtained are to facilitate the process of quadtree (QT) decomposition later in segmentation stage. Besides, all images are resize into 4096*4096. The pixel value of CXR image is a number that represents the brightness of the pixel and stored as an 8-bit integer which gives a range from 0-255. Since the pixel value of the grey-scale image is a single number, therefore, it is simpler

compared to RGB image where its pixel value is a combination of red, green and blue. After that, all the grey-scaled images will be converted into another representation which is QT. The reason of choosing the QT representation for feature extraction instead of images is because the computational time required for processing an grid image is longer than the processing of QT representation and the QT representation also able to accurately maintain the structural information of the region of interest of an image which means that it might be helpful in image classification. The QT representation is shown below.

Quadtree (QT) is a tree data structure in where each internal node has exactly four children. QT is two dimensional analogy of octrees and is most often used to partition a two-dimensional space by recursively subdividing it into four quadrants or regions. The data associated with a leaf cell varies by application, but the leaf cell represents a "Unit of interesting spatial information".

Every parent node in QT only consists of four child nodes which also known as quadrants and each node is representing the information of a certain block on the image such as mean colour pixel value, standard deviation, and the others. Each node will be tested with a criterion of homogeneity. If the homogeneity criterion is satisfied, there will be no further subdivision of that block. Otherwise, the blocks must be subdivided into quadrants, sub-quadrants and so on. The decomposition process is repeated until there is no further subdivision is required. One of the advantages of applying the image manipulation is that the set operations of union and intersection can be done simply and quickly. CXR in this research is decomposed into four equal sized blocks – quadrants
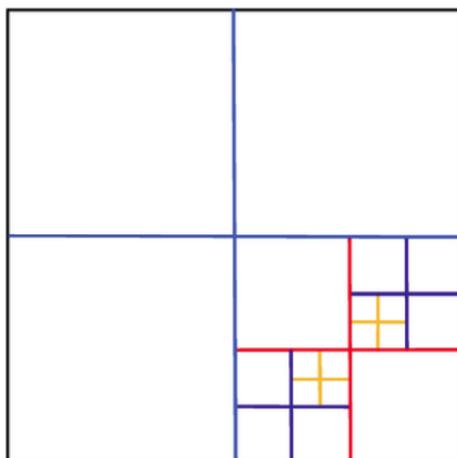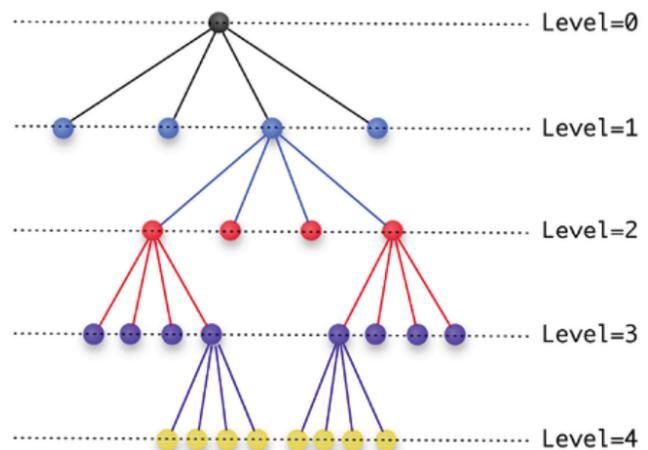

Figure 2: Quadtree Representation.

and then test each of them with homogeneity criterion. Decomposition of blocks will stop when criterion of homogeneity is met, otherwise it will continue divide iteratively into sub-quadrants that are more homogeneous until the test criterion met which resulting in a different size of squared blocks.

4.3. Feature Extraction

Feature extraction can be considered as finding out a group of the most effective feature for classification; that is compressing from high dimensional feature space to low dimensional feature space, in other words it is the method to extract the most representative information from raw data. It is essential to note that choosing a feature extraction method is also an important factor in achieving good classification accuracy. In this research, features that able to represent the characteristic of tuberculosis on the images which helps in the process of detection of mycobacterium will be extracted from quadtree as it is representing the CXR images. Therefore, the extracted features that are significantly distinguishable between normal and TB detected CXR images will be chosen and applied as the input to the classifier for image classification to rise the result of classification. In this research, features that are chosen to extract is texture, colour and shape. All features is extracted from the 4096*4096 images and quadtree representation images for comparison in later stage.

4.4. Image Classification

Image classification is the process that classifies the input images into different classes which is class normal and class TB detected based on extracted information from feature extraction stage. In this research, three different type of classification will be applied. First classification method will use multilayer perception (MLP), K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) as standalone classifier respectively. Second classification method will used Convolutional Neural Network (CNN) as classifier and third method will use Ensemble classifier on which is the voting method on first and second classification method.

4.4.1. Training and Validation

Number of images used for training and validation from total of 800 images will be randomly distributed in the ratio of 90% and 10% respectively. 720 CXR images will be used for training model and 80 CXR images will be used for validation purpose.

4.4.2. Result and Measurement

Result measurement is a method to determine the performance of the algorithm proposed based on the result obtained in this research. Accuracy and Sensitivity will be calculated in this research.

Sensitivity is defined as the ratio of the correctly classified TB cases to the total number of TB cases.

$$Sensitivity\ (\%) = \frac{TP}{TP + FN} \times 100$$

Where,

TP refer to True Positive,

FN refer to False Negative.

Accuracy is defined as the ratio of number of correct classification of input – output data to the total number of testing data.

$$Accuracy\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Where,

TP refers to True Positive,

FP refers to False Positive,

TN refers to True Negative,

FN refers to False Negative.

All the CXR images have been examined and labeled properly by the expert. Therefore, the result obtained by using the proposed approach can be used to compare with the class label of the respective TB image set in the dataset.

5. CONCLUSION

Overall this proposed approach aims to improve the accuracy of detecting tuberculosis. Beside Chest X-Ray datasets that were proposed to be used in this paper, Chest X-Ray (CXR) datasets from different regions in the world such as CXR from Africa or from western countries can be applied too. Furthermore, the total number of datasets used as training and validation suggested to increase as it might improve overall performance. Last but not least, features that are not mentioned in this paper are also applicable for the feature extraction process.

REFERENCES

[1]   Mathews A, A VR. Detection of Tuberculosis Using Chest X-Rays. International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) 2015; 4(6).

[2]   Antony B, NB K. Lung Tuberculosis Detection Using X-Ray Images. International Journal of Applied Engineering Research 2017; ISSN, 15196-15201.

[3]   Melendez J, Ginneken BV, Maduskar P, Philipsen RH, Reither K, Breuninger M, Sánchez CI. A Novel Multiple-Instance Learning-Based Approach to Computer-Aided Detection of Tuberculosis on Chest X-Rays. IEEE Transactions on Medical Imaging 2015; 34(1): 179-192.
https://doi.org/10.1109/TMI.2014.2350539

[4]   Rusdah, Winarko, E., Wardoyo, R. Preliminary diagnosis of pulmonary tuberculosis using ensemble method. International Conference on Data and Software Engineering (ICoDSE) – IEEE 2015.
https://doi.org/10.1109/ICODSE.2015.7436993

[5]   Lee J, Lee S-K, Yang S-I. An Ensemble Method of CNN Models for Object Detection. 2018 International Conference on Information and Communication Technology Convergence (ICTC) 2018.
https://doi.org/10.1109/ICTC.2018.8539396

[6]   Li Y, Meng MQ-H, Li S, Chen W, You Z, Guo Q, Liang H. A quadtree based neural network approach to real-time path planning. 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO) 2007; 1350-1354.

[7]   He Y, Xu C, Khanna N, Boushey CJ, Delp EJ. Food image analysis: Segmentation, identification and weight estimation. 2013 IEEE International Conference on Multimedia and Expo (ICME) 2013; 2744-2748.
https://doi.org/10.1109/ICME.2013.6607548

[8]   Treebupachatsakul T, Poomrittigul S. Bacteria Classification using Image Processing and Deep learning. 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) 2019; 1-3.
https://doi.org/10.1109/ITC-CSCC.2019.8793320

[9]   Paras Lakhani BS. Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. THORACIC IMAGING: Deep Learning at Chest Radiography 2017.
https://doi.org/10.1148/radiol.2017162326

[10] Liu C, Cao Y, Alcantara M, Liu B, Brunette M, Peinado J, Curioso W. Detecting tuberculosis in chest X-ray images using convolutional neural network. IEEE International Conference on Image Processing (ICIP) 2017.
https://doi.org/10.1109/ICIP.2017.8296695

[11] Marquez GR, Escalante HJ, Sucar LE. Simplified Quadtree Image Segmentation for Image Annotation. Chapter in Proceedings of the 2010 Automatic Image Annotation and Retrieval Workshop 2010; 719: 24-34.

[12] Deivanayagampillai N. Statistical based image comparision by quad tree segmentation. International Journal of Research in Engineering and Applied Sciences (IJREAS) 2017; 7(1): 57-63.