

A statistical approach on pulmonary tuberculosis detection system based on X-ray image

Ratnasari Nur Rohmah¹, Bana Handaga², Nurokhim³, Indah Soesanti⁴

^{1,2}Universitas Muhammadiyah Surakarta, Jalan A. Yani Kartasura, Surakarta, 0271-717417, Indonesia

³Pusat Teknologi Keselamatan dan Metrologi Radiasi, BATAN, Indonesia

⁴University of Gadjah Mada, Indonesia

*Corresponding author, e-mail: rnr217@ums.ac.id¹, nurokhim@batan.go.id

Abstract

This paper presented the research result on the design of pulmonary TB (Tuberculosis) detection systems using a statistical approach. The study aimed to address two problems in detecting pulmonary TB by doctors, especially in remote areas of Indonesia, namely the long waiting time for patients to get the doctor's diagnosis and the doctor's subjectivity. We used hundreds of X-ray images from radiology department of Sardjito Hospital, Yogyakarta, as primary data and thirty data from various sources on the internet as secondary data. Using statistical approach, we exploited statistical image feature from image histogram, examined two statistical methods of PCA and LDA transformation for feature extraction, and two minimum distance classifier in image classification. We also used histogram equalization in the image enhancement process and bicubic interpolation in image segmentation and template making. Test results on primary and secondary data images show the identification accuracy of 94% and 83.3%, respectively.

Keywords: detection, pulmonary tuberculosis, statistical approach

Copyright © 2019 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

The X-ray images have been used to support the diagnosis of pulmonary TB (Tuberculosis) disease [1-4]. The Ministry of Health Decree No. 364/Menkes/SK/V/2009 states that examination on sputum and patient's lung X-ray images should be done to diagnose Tuberculosis. However, there are problems with the low ratio of the radiologist to the number of patients and the subjectivity factor in radiologist's diagnosis. The low ratio makes patient must wait for a long time to receive the results on X-ray image reading, and the subjectivity factor influences the diagnosis that based on radiologist's visual observation on X-ray lung images. The observation of X-ray images by one radiologist to diagnose pulmonary TB may be interpreted differently by other radiologists.

This study aimed to address those two problems by designing a system for pulmonary TB detection based on X-ray image using computer assistance [1, 2, 5-7]. Not only this automatic TB detection would reduce the patient's waiting time to get the result on radiologist's diagnosis, but also the quantitative calculation based on computer algorithm would reduce the subjectivity factor. We proposed a pattern recognition system to identify lung X-ray images as TB images or non-TB images [1, 2].

Several studies of the detection of pulmonary TB based on X-rays images have been conducted by various researchers [1, 2, 8-15]. Each researcher used their particular dataset and get various TB detection accuracy result, such as 69.8-81.6% [8], 90.3% [10], 78.3% [11], 84% [12], and 72.8% [15]. Some works involve the CBIR (content-based image retrieval) method to find the five most similar ROI models and then using 'graph cut' area method to get ROI [13, 14]. Other researcher proposed methods that select the TB images based on the specific radiological features of pulmonary TB, which is the lung cavity [15]. After the cavity models templates are made, the study conducted a coarse identification of cavity, contour segmentation, and fine identification on the cavity [15]. Then, it used the support vector machine (SVM) as the classifier to classified images as TB or non-TB images. Nevertheless, all these previous research show an opportunity for other researcher to improve the accuracy results, as well as to find more efficient systems. In this paper we proposed a statistical approach TB

detection system that has shorter but effective process. The system design consisted of object locator design to define the region of interest (ROI) image, feature image extraction method design, classifier design and training, and system performance evaluation [1, 2] as shown in Figure 1.

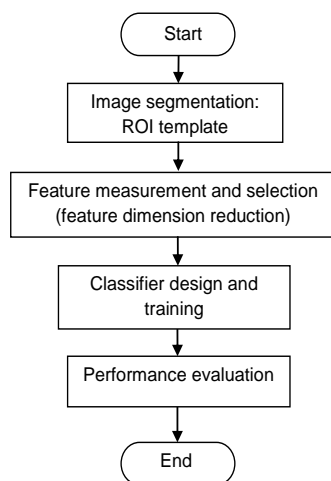


Figure 1. Flowchart of pulmonary TB detection design

The X-ray images used in diagnosis is always in greyscale so that the textural features become the dominant features of this data images. The texture is essential in the characterization of tissue and recognition of pathological structure [16-18]. Such statistical textural features produce an insignificant number of features that relevant to distinguish the textural condition of images [19]. However, the statistical textural features of the images may be redundant. Thus, the feature's image dimension reduction is needed to obtain less number but mutually exclusive features, which may represent the most important characteristic of the textural features [19, 20].

Image features extraction was conducted on ROI image, which was the lung area image. It is important to determine ROI properly so that the feature extraction process gives significant features to be used in image classification [21]. This feature must be reliable, which requires the same consistency of ROI in each image. However, not all X-ray thorax images have exactly the same body size and position; thus, it is difficult to obtain similar ROI. In addition, there are non-uniformity of image brightness and contrast as well. An ROI capture method and image enhancement method are needed to address those problems. Those methods must be an appropriate one for image classification purposes; hence, the results should support the class separation between groups of data.

The proposed design was a method for pulmonary TB detection based on the first-order statistical textural feature of X-ray image. The image segmentation in this study was designed to be shorter than the segmentation process in the previous studies [1, 2, 8-15], by using predefined ROI template and image resizing to match with template size. In quality image enhancement, we used a statistical-approach by using histogram equalization. We also employed feature dimensions reduction in image feature extraction. It will select the most significant image feature to be used in the classification process. Statistical theoretical-decision approach is employed in classifier design. In this approach, a classification was made by using a 'decision function' or 'discriminator function'. The discriminator function is a function that indicates the relationship between variables that separate the data classes [1, 2, 22]. The complete design is expected to detect pulmonary TB in shorter calculation and more accurate than those previous methods [8-15].

2. Research Method

The data used in this study consisted of primary and secondary data. Primary data were digital X-ray greyscale images in .bmp format in various sizes. These data were taken in the

radiology department of Sardjito Hospital, Yogyakarta. Validation of primary data was based on the results of the doctor's diagnosis, stated in medical record accompanied the image data. Secondary data in this research were images obtained from various sources on the internet, in the formats of .jpg and .png, and consisted of not only greyscale data but also RGB images. We used 25 of normal (non-TB) primary images and 25 of TB primary data images as reference images in the classifier design and in training process. We also used other 50 images of primary data and 30 secondary data to evaluate the system performance.

2.1. Image segmentation Design–ROI Template Making

Image segmentation in TB detection in this research was designed using ROI template; therefore, defining the optimum ROI template was essential. The template making steps is shown in Figure 2. Normal reference image data was used as input in the ROI template making process. This process included images cropping, resizing, image averaging, and grey level thresholding technique. The process of the image averaging was done to address non-uniformity in size and body position. Twenty five non-TB images were cropped to get images with semi exclusive lung area. This process then followed by image size equalization (resizing image) with bicubic interpolation. The average image was then the sum of all resized digital images reference divided by 25.

As mention above, we used bicubic interpolation in image rezing. Bicubic interpolation is an interpolation used in image spatial transformation, which defines the geometric relationship between input and output images. The output image was a spatial manipulation of the input image where points in pixel coordinate were different than the input image. The change of coordinate points required mapping of the new pixel value (output image) in such a way so as not to affect the image quality. Bicubic interpolation used the intensity value of the 4×4 neighbor pixels around the new pixel to defined pixel value of any particular new pixel.

We applied thresholding technique to average image in the final process of template making. Gray level thresholding technique is an algorithm that divides image area based on the similarity of pixel values to a certain criterion setting [22]. Since the average image had two histogram peaks (bimodal) as shown in Figure 3, the thresholding method was the suitable method in this image segmentation design [22]. We made four ROI templates using four threshold levels around the Otsu's level of 120, 128, 136 and 144. These levels were selected based on visual observation on average image histogram. We also made five different sizes of ROI templates to investigate the effect of bicubic interpolation on the textural image feature.

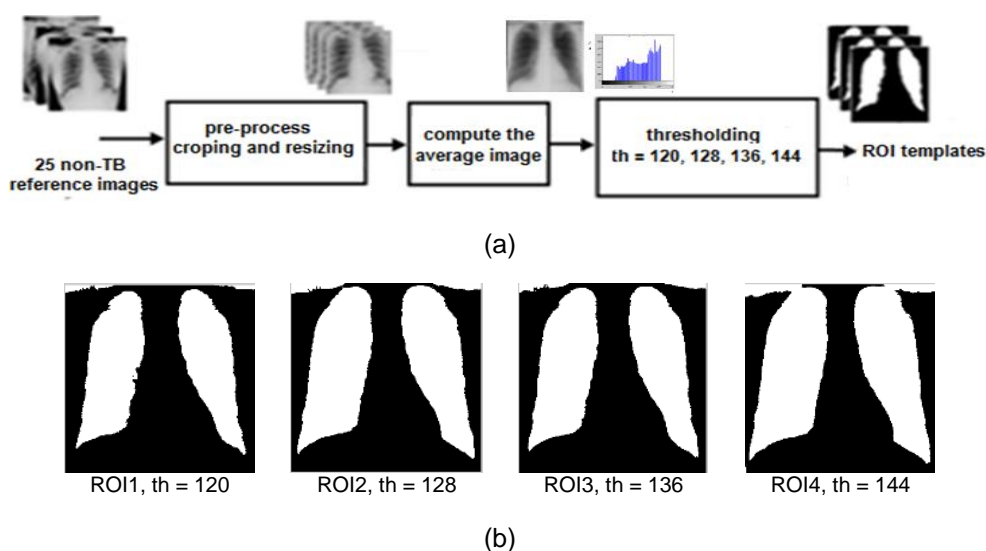


Figure 2. (a) Block diagram of ROI templates making; (b) ROI templates result

2.2. Image Feature Extraction–feature Measurement and Selection

Since most of the medical image is represented in gray level, the texture becomes an important characteristic of the image. One of the feature extraction techniques for the textural

feature is the first-order statistical method by measuring the characteristics of image histogram [23]. In this study, feature calculation was preceded by a histogram equalization process to overcome non-uniformity of images quality. We calculated five statistical characteristics of image histogram: mean, standard deviation (STD), skewness, kurtosis, and entropy. Using these five features, we selected one most important feature using feature dimension reduction methods, as in (1). We compared two feature dimension reduction methods, namely principal component analysis (PCA) and linear discriminant analysis (LDA), to find the best one.

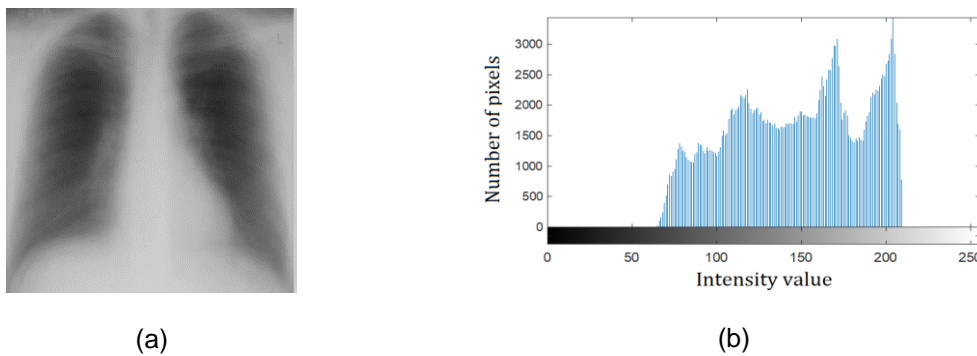


Figure 3. (a) The average image of 25 reference images and (b) its bimodal image histogram

PCA aims to maximize between-class data separation while LDA not only tries to maximize between-class data separation but also minimize within class data separation [24-30]. PCA optimizes the transformation matrix by finding the largest variations in the characteristics of space origin. On the other hand, LDA seeks the largest ratio of variants between two classes and inter-class variants to project the origin feature space into the sub-space, as in (2) to (6). We compared those two methods in classifier design.

The PCA algorithm was:

- Calculate centered feature matrix
- Calculate the Eigenvector and Eigenvalue of the centered feature matrix
- Choose the Eigenvector with the greatest Eigenvalue for matrix dimension reduction
- Reduce the dimension of centered feature matrix, transforming centered feature matrix using selected Eigenvector:

$$\text{Feature vector} = [\text{Eigenvector}]^T \times [\text{centered feature matrix}] \quad (1)$$

Meanwhile, the LDA algorithm consisted of:

- Calculation of the scatter matrix within five measured features:

$$S_w = \sum S_i \quad (2)$$

with: S_i for each feature:

$$S_i = \sum (x - m_i)(x - m_i)^T \quad (3)$$

Where,

m_i : mean of an n-sample of each feature

$i = 1, 2, 3, 4, 5$ (five features were calculated in this research)

- Calculation the scatter matrix between five features measured:

$$S_B = \sum n_i (m_i - m)(m_i - m)^T \quad (4)$$

where”

m_i : mean of an n-sample of each feature

m : mean of an all-sample of five featured

n_i : number of samples of each five measured features

- Calculation of Eigenvalue and Eigenvector from:

$$S_B \times [Eigenvector] = [Eigenvalue] \times S_w \times [Eigenvector] \quad (5)$$

- Dimension reduction of centered feature matrix by transforming feature matrix using selected Eigenvector:

$$Feature\ vector = [Eigenvector]^T \times [feature\ matrix] \quad (6)$$

2.3. Classifier Design

Classifier to identify lung image as TB or non-TB image was designed based on statistical feature selection. The discriminator function in this research was made by calculating the features distance of the class members to the average feature of each class, as in (7). Two types of statistical approach classifiers, namely Euclidean distance classifier and Mahalanobis distance classifier was employed in this design. If the Euclidean distance was used, as in (8) and (9), the classifier function was called a minimum Euclidean distance classifier. However, if Mahalanobis distance was used, as (10) and (11), the classifier function was called as a Mahalanobis distance classifier. We used these two classifier methods to examine which method was better to be used in this study, in accordance with the possibility of difference distribution of data sample between the two classes. Minimum distance classification function was:

$$D(x)_{12} = D(x)_1 - D(x)_2 \quad (7)$$

and the decision is:

if $D(x)_{12} < 0$ the image is a non-TB lung image

else $D(x)_{12} > 0$ the image is a TB lung image

with $D(x)_1$ is a distance between any tested images' feature to feature of non-TB reference class, and $D(x)_2$ is a distance between any tested images' feature to feature of the TB-image reference class. For Euclidean distance classifier, the distance was calculated by:

$$D(x)_1 = \sqrt{(x - \mu_1)^2} \quad (8)$$

and

$$D(x)_2 = \sqrt{(x - \mu_2)^2} \quad (9)$$

where:

x = a feature of the tested image

μ_i = mean feature of all feature images from i -class (this study used 2 classes, TB class or non-TB class) of the reference image

In Mahalanobis distance, not only data means used in distance calculation but also the variant of each class (covariance if there were multiple data in each class). In Mahalanobis distance classifier, the distance was calculated by:

$$D(x)_1 = \sqrt{(x - \mu_1)^T \cdot (C_1)^{-1} \cdot (x - \mu_1)} \quad (10)$$

$$D(x)_2 = \sqrt{(x - \mu_2)^T \cdot (C_2)^{-1} \cdot (x - \mu_2)} \quad (11)$$

where:

x = a feature of the tested image

μ_i = mean feature of all feature images from i -class (TB class or non-TB class) of reference image

C_i = Covariance image features from i -class (TB class or non-TB class) of the reference image

3. Results and Analysis

3.1. Result and Analysis on Systems Design

The used of ROI template showed a significant effect on determining ROI properly so that the feature extraction process gave significant features to be used in image classification.

Figure 4 shows how mean histogram feature calculation on segmented images gave better result in cluster separation of reference image rather than on unsegmented images. We found the same tendencies on four other statistical image features, std, skewness, kurtosis, and entropy. Selection on ROI template was performed based on its contribution in image reference class separation based on image feature. The image feature was the result of a reference image feature extraction using PCA transformation. All templates were tried in image segmentation process before image feature extraction.

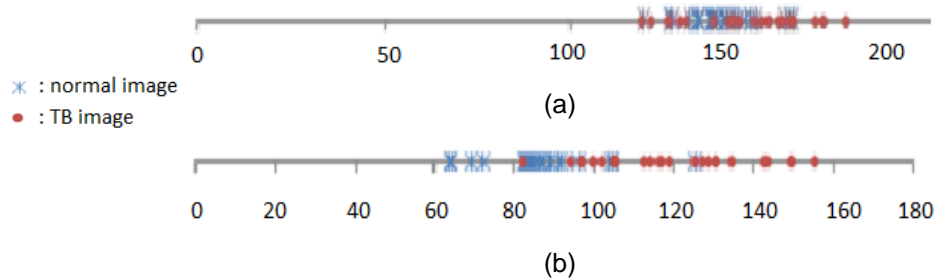


Figure 4. Cluster separation of image reference (normal image and TB image) based on statistical image feature: 'mean'; (a) features calculated from unsegmented images and (b) features calculated from ROI image using ROI template

We evaluated of various ROI templates shape (ROI1, ROI2, ROI3, and ROI4) and various ROI size (2048x2048, 1024x1024, 512x512, 256x256, and 128x128 pixels). The results are shown in Table 1 and in Figure 5. The results show that the change in shape of the ROI template (ROI1, ROI2, ROI3, and ROI4), affect the distance class of image clustering. The ROI3 template had the optimum distance in image clustering based on Mahalanobis distance. The results also show that the changes in size did not affect the distance as shown in Figure 5. The bicubic interpolation used in resizing image did not change image histogram significantly. These results were validated by comparing the image histogram of actual and resized images as shown in Figure 6.

Table 1. Comparison of Distance between TB Class and Non-TB Class of Reference Imageclusteringonvarious ROI Templates Test

Distance	ROI1	ROI2	ROI3	ROI4
Euclidean	30.223	28.179	25.925	23.233
Mahalanobis of data mean from TB class to normal class	30.6480	33.856	35.038	34.156
Mahalanobis of data mean from normal class to TB class	8.5534	8.8567	9.0658	8.9159

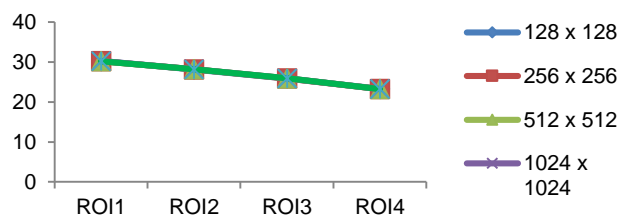


Figure 5. The distance comparison between TB class and normal class of reference image clustering on various ROI size test

Evaluation of LDA and PCA as the feature selection method was conducted by comparing the effect of each method on reference image clustering. In this evaluation, image segmentation was done using the ROI3 template (the selected template). Table 2 shows the evaluation results. The PCA transformation was more superior in separating the two image

classes. These results indicated that the PCA transformation was more appropriate for feature extraction in this study.

Table 2 also displays the comparison between the Euclidean distance and Mahalanobis distance of image clustering. The results show that the Mahalanobis distance was more appropriate for this study. In the Mahalanobis distance, not only data means was used in distance calculation but also the variant of each class. The results show that the distribution of data sample between the two classes was not the same. Based on this result, we selected the Mahalanobis distance to define the discriminate function in the classifier method design.

3.2. Results and Analysis Systems Performance Test

Fifty test images of primary data used in the performance evaluation. Those were 20 images of pulmonary TB, 20 healthy lung images, and 10 images with bronchitis. In addition to the primary data, 30 secondary data were also used in the experiment. They were 15 images of pulmonary TB, 10 healthy lung images, and 5 images with bronchitis. Performance evaluation was done by looking at the accuracy of the identification. The accuracy in this study was measured by observing the comparison between the test results from a computer and the doctor's diagnosis. The accuracy was measured by calculating parameters of accuracy, FAR (false acceptance ratio), and FRR (false rejection ratio).

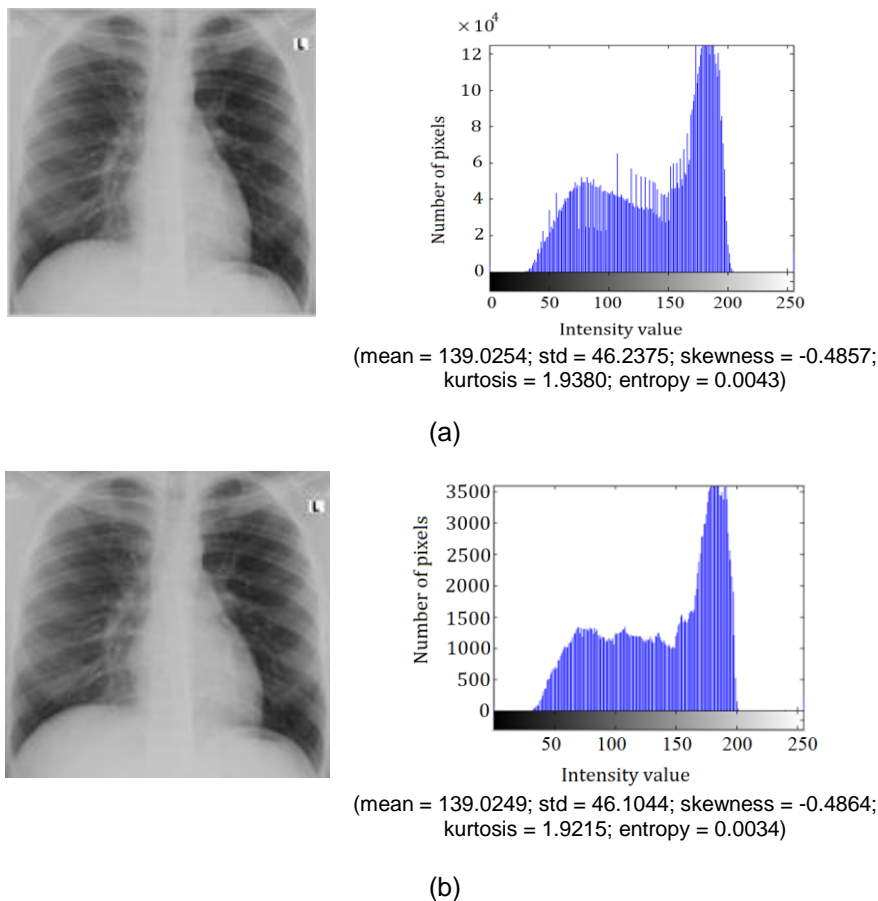


Figure 6. The comparison of statistical image features between (a) original size image (2864x3040 pixel) and (b) resized image (500x500 pixel), their histograms, and their calculated five statistical features

Image identification was conducted in processes as shown in Figure 7. It consist of: image segmentation used the ROI3 template, PCA feature selection method, and Mahalanobis distance classifier. Test results on primary data image show that the identification accuracy in this study was 94%. Furthermore, the results show that there were 9.5% false identification

where non-TB images were assessed as TB images and 3.5% false identification where TB images were assessed as non-TB images.

The experiment with the different size of ROI templates; 128×128 pixels and 2048×2048 pixels showed the same accuracy. Image size matching with the ROI template on this study also was done by resizing the image with bicubic interpolation method. The same accuracy result for both 128×128 pixels and 2048×2048 pixels of ROI templates showed that this interpolation had no significant effect on statistical histogram features. However, these difference in size affected in the computing time by approximately 2 seconds. The time required for the identification after manual cropping of a single image with 128×128 pixel resizing was approximately 1 second, while the 2048×2048 pixels images resizing was approximately took 3 seconds.

Table 2. Euclidean and Mahalanobis Distance between Normal and TB Images of Image Clustering Based on Image Feature in PCA and LDA Transformation

Distance	PCA	LDA
Euclidean distance between two classes	25.9254	0.5649
Mahalanobis distance of data mean from TB class to all data from normal class	35.0378	27.351
Mahalanobis distance of data mean from the normal class to all data from TB class	9.0658	12.6099

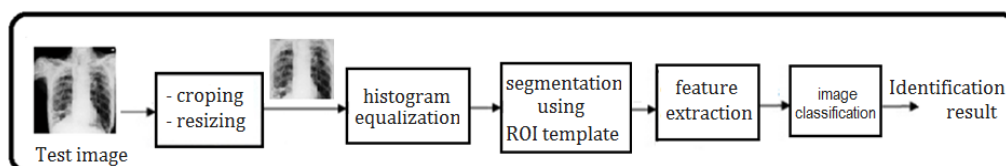


Figure 7. Steps in TB identification of a test image

Test results of secondary data images show that the TB detection accuracy in this study was 83.3%. Furthermore, the results show that there were 18.8% false identification where non-TB images were assessed as TB images and 14.3% false identification where TB images were selected as non-TB images. The secondary data accuracy was lower than the primary data, which was influenced by:

- The format of the secondary data images was in a .jpg and .jpeg while the primary data were in .bmp format.
- Secondary data image quality was not as good as the primary data. This was caused by the image acquisition process. The secondary data were downloaded from the internet while the primary data were obtained from the digital X-ray images at the hospital.
- The secondary data had more size variation than the primary data.
- The validation was based only on the information that accompanied the image.

4. Conclusion and Future Studies

The method of pulmonary TB detection using computer assistance proposed in this research was based on X-ray images features using a statistical approach. It identifies TB or non-TB images using ROI template segmentation process, feature extraction using PCA transformation, and Mahalanobis distance classifier. The method has a shorter process than the other existing methods and it performed excellently in our particular dataset. Test results on primary data image show that the identification accuracy in this study was 94%, with FAR: 9.5% and FRR: 3.5%. The results also show that the use of bicubic interpolation for image resizing has no significant effect on statistical histogram features. In the other hand, the shape of the ROI image was a significant parameter to obtain the best image identification result. The proper ROI image produced the best-measured image feature to be used in the classification process. The systems proposed only classify input data image as TB and non-TB, and not classify the severity level of TB image. This results gives opportunity for further study in determining specific feature to define the level of TB stage.

References

- [1] Rohmah RN, Susanto A, Soesanti I, Tjokronagoro M. *Computer Aided Diagnosis for lung tuberculosis identification based on thoracic X-ray*. Information Technology and Electrical Engineering (ICITEE), 2013 International Conference. 2013: 73-78.
- [2] Rohmah RN, Susanto A, Soesanti I. *Lung tuberculosis identification based on statistical feature of thoracic X-ray*. Proceeding of The 13th International Conference on QiR. Yogyakarta. 2013: 19-26.
- [3] Miller FJW. *Tuberculosis in Children Evolution, Epidemiology, Treatment, Prevention*. New York: Churchill Livingstone Inc. 1982.
- [4] Puspongoro HD, et. al. *Standar Pelayanan Medis Kesehatan Anak*. Edisi I. Badan Penerbit IDAI. 2004.
- [5] Kanazawa K, Kawata Y, Niki N, Satoh H, Ohmatsu H, Kakinuma R, Kaneko M, Moriyama N, Eguchi K. Computer-aided diagnosis for pulmonary nodules based on helical CT images. *Computerized medical imaging and graphics*. 1998; 22(2): 157-167.
- [6] Castleman KR. *Digital Image Processing*. 2nd Edition. New York: CRC Press. 1996.
- [7] Qin C, et.al. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical engineering online*. 2018; 17:113. <https://doi.org/10.1186/s12938-018-0544-y>.
- [8] Rohilla A, et.al. *TB Detection in Chest Radiograph Using Deep Learning Architecture*. Proceeding of 5th International conference on Emerging Trends in Engineering, Technology, Science and Management (ICETETSM -17). 2017: 136147.
- [9] Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017. 284(2): 574–582.
- [10] Hwang S, Kim HE, Jeong J. A novel approach for tuberculosis screening based on deep convolutional neural networks. *Medical imaging 2016: computer-aided diagnosis*. 2016; 9785: 97852W. <https://doi.org/10.1117/12.2216198>. 91.
- [11] Jaeger S, et.al. *Detecting Tuberculosis in Radiographs Using Combined Lung Masks*. Proceeding of 34th Annual International Conference of the IEEE EMBS. 2012: 4978–4981.
- [12] Jaeger S, et.al. Automatic Screening for Tuberculosis in Chest Radiographs: a survey. *Quantitative Imaging in Medicine and Surgery*. 2013; 3(2): 88–99.
- [13] Jaeger S, et.al. Automatic Tuberculosis Screening using Chest Radiographs. *IEEE Transactions on Medical Imaging*. 2014; 33(2): 233–245.
- [14] Candemir S, et.al. Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration. *IEEE Transactions on Medical Imaging*. 2014; 33(2): 577–590.
- [15] Xu T, et.al. Novel coarse-to-fine dual scale technique for Tuberculosis cavity detection in chest radiographs. *EURASIP Journal on Image and Video Processing*. 2013; 3: 1-18.
- [16] Mitrea D, Nedeveschi S, Lupsor M, Badea R. *Exploring the Textural Parameters obtained from Ultrasound Images for Modeling the Liver Pathological Stages in the Evolution towards Hepatocellular Carcinoma*. IEEE Conference of Automation. Quality and Testing. Robotics. 2008; 3: 128-133.
- [17] Kumar SS, Moni RS, Rajeesh J. *Liver tumor diagnosis by gray level and contourlet coefficients texture analysis*. Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference. 2012: 557-562.
- [18] Veenland JF, Grashuis JL, Weinans H, Ding M, Vrooman HA. Suitability of texture features to assess changes in trabecular bone architecture. *Pattern Recognition Letters*. 2002; 23(4): 395–403.
- [19] Aggarwal N, Agrawal RK. First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing*. 2012; 3(02): 146-153.
- [20] Nurhayati OD, Susanto A, Widodo TS, Tjokronagoro M, Principle Component Analysis combined with First Order Statistical Method for Breast Thermal Images classification. *IJCST*. 2011; 2(2): 12-18.
- [21] Singh S, Kumar V. *SVM Based System for classification of Microcalcifications in Digital Mammograms*. Proceeding of the 28th IEEE EMBS Annual International Conference 2006: 4747–4750.
- [22] Gonzalez RC, Woods RE. *Digital Image Processing*. 3rd Ed. New Jersey: Pearson Education. 2008.
- [23] Jain AK. *Fundamentals of Digital Image Processing*. NJ: Prentice- Hall Inc. 1989.
- [24] Mazanec J. et.al. Support Vector Machines, PCA, and LDA in Face Recognition. *Journal of Electrical Engineering*. 2008; 59(4): 203–209.
- [25] Abd-Almageed W, Davis L. *Human detection using iterative feature selection and logistic principal component analysis*. Robotics and Automation, 2008, ICRA 2008. 2008: 1691-1697.
- [26] Lei J, et.al. An Image Reconstruction Algorithm for Electrical Capacitance Tomography Based on Robust Principle Component Analysis. *Sensors*. 2013; 13: 2076-2092.
- [27] Partridge M, Calvo RA. Fast dimensionality reduction and simple PCA. *Intelligent Data Analysis*. 1998; 2(3): 203-214.
- [28] Ranjith M, Balaji RM, Surjith KM, Dhyaneswaran J, Baskar A. *Content based Image Retrieval for Medical Image (cerebrum intract) using PCA*. Conference Proceedings RTCSP. 2009: 124–126.
- [29] Sinha U, Kangarloo H. Principal component analysis for content-based image retrieval. *Radiographics*. 2002; 22(5): 1271-1289.
- [30] Zhou J, Jin Z, Yang J. *Multiscale saliency detection using principle component analysis*. Neural Networks (IJCNN), The 2012 International Joint Conference. Brisbane. 2012: 10-15, 2012.