Analysis of Tuberculosis (TB) on X-ray Image Using SURF Feature Extraction and the K-Nearest Neighbor (KNN) Classification Method

Reyhan Achmad Rizal[1)], Nurlela Octavia Purba[2)], Lidya Aprilla Siregar[3)], Kristina Sinaga[4)], Nur Azizah[5)]

[1,2,3,4,5] *Faculty of Technology and Computer Science, Universitas Prima Indonesia, Medan, Indonesia*

Abstract— With current technological developments, machine learning has become one of the most popular methods, one of the popular machine learning algorithms is k-nearest neighbors (KNN). Machine learning has been widely used in the medical field to analyze medical datasets, in this study the k-nearest neighbors (KNN) machine learning algorithm will be used because of its good level of accuracy in recognition and is included in the supervised learning algorithm group. The results showed the k-nearest neighbors (KNN) method in recognizing x-ray images of tuberculosis (TB) using SURF feature extraction with an average accuracy of 73%.

Keywords—*Tuberkulosis, KNN, SURF.*

1. Introduction

Tuberculosis (TB) is still a major public health problem in many countries that can cause death [1]. During the last decade, the number of patients infected with tuberculosis has been increasing every year [2]. Tuberculosis is a bacterial infection in the air caused by Mycobacterium tuberculosis (MTB) which attacks any part of the body and is generally the lungs, the signs of tuberculosis are: cough often lasts longer than 3 weeks with or without sputum production, coughing up blood, chest pain, loss of appetite, unexpected weight loss, night sweats, fever and fatigue [3]. Tuberculosis is transmitted through the air when a person coughs up the lungs [4]. Rapid and accurate diagnosis of tuberculosis is essential to ensure proper therapy and treatment [5].

Research in the field of bioinformatics is currently becoming popular as a solution for the medical world [6], with the development of technology in the medical field today, resulting in machine learning being one of the most popular methods where with the help of this system it is possible to make a misdiagnosis by medical experts. can be avoided. [7]. Some of the popular machine learning algorithms are K-Nearest Neighbor (KNN), support vector machine (SVM) [8], Naive Bayes, Radial Basis Function (RBF), [9] LDA and Learning Vector Quantization (LVQ). Machine learning has been widely used in the medical field to analyze medical datasets, [10] to analyze cancer gene datasets in patients using supervised machine learning algorithms to classify cancer cells based on microRNA gene expression. [11] analyzed the breast cancer dataset using the support vector machine (SVM) algorithm, the SVM algorithm in ML is used to look for a pattern of data from a set of data that can produce predictions to determine whether live breast cancer cells are malignant or benign. [12] analyzed the breast cancer dataset using a sequential minimal optimization algorithm.

in this study, the K-Nearest Neighbor (KNN) machine learning algorithm will be used in classifying tuberculosis (TB) because K-Nearest Neighbor (KNN) has several advantages, namely that it is tough on noisy training data and is effective when processed on large training data. 13]. The x-ray image files used for training and testing were taken from http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip with a total sampling of 1000 images of tuberculosis (TB), there are several studies related to implementing K-Nearest Neighbor (KNN) algorithm and SURF feature extraction: [14] Classify Hernia Disc disease and Spondylolisthesis of the spine. The data were arranged in two different but related tasks, namely the “normal” and “abnormal” categories. The results showed that the test of the K-NN classifier was 83%. The average length of time required to classify the K-NN classifier is 0.000212303 seconds. [13] analyzed lung disease using the K-NN algorithm at the Aloe Saboe hospital, Akota Gorontalo, the results of the predictions carried out by the K-NN algorithm yielded a fairly high accuracy reaching 91.90%. [15] used the K-Nearest Neighbor (KNN) algorithm to predict patients with diabetes at many public health centers in Bulukumba district, the results of the accuracy obtained from the test were 68.30%. [16] compared to other models such as Naïve Bayes, Support Vector Engine, K-Nearest Neighbor and artificial neural networks. [17] used the KNN method to compare with normalized euclidean distances, Manhattan and normalized Manhattan to achieve optimization results or optimal values in finding the closest neighbor distance. [18] Comparing the feature extraction of SURF and HOG with the percentage of accuracy of SURF 85.83% while the accuracy of HOG extraction is 83.50%. [19] in his research using SURF and SIFT feature extraction to obtain the characteristics of a person's feet where the test results have been obtained, then SURF feature extraction is better used than SIFT feature extraction in an individual identification system using foot

images based on accuracy in grouping images. [20] using GLCM and classifying it using K-KNN research shows that the GLCM and KNN methods are able to classify bougainvillea plants based on leaf texture with an accuracy of 87% on the input of adherence values K = 3, K = 5, K = 7 and K = 9 while accuracy The lowest is in the input neighbor value K = 1, which is 75%.

2. Methods of Research

2.1. Dataset

The data used in this study are images with a size of 256x256. Image files used for training and testing were taken from http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip with a total sampling of 622 tuberculosis (TB) disease images.

2.2 Research Steps

The general research steps built in this study can be seen in Figure 1 below :

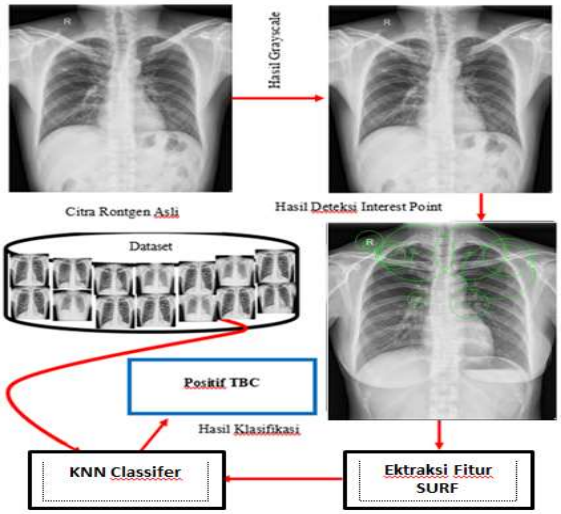

Figure 1. General research steps

Figure 1 describes the research scheme in general, in this scheme, there are two processes, namely the training process and the testing process, the color input image training process will be converted into a gray image to save computation time from three color channels to 1 color channel, then extracted using SURF and the extraction results will be saved as a pattern model. Meanwhile, the testing process goes through the same stages as the stages that are passed during training only when the classification stage uses the K-Nearest Neighbor (KNN) method.

3.2 Testing

3. Results and Discussion

The sample of tuberculosis (TB) x-ray images used in this study consisted of 622 TB images, with 326 negative TB images and 296 positive TB images. Tuberculosis (TB) image samples used as training data amounted to 60% of the 622 images and 40% were used as testing data. The data used in this study are images with a size of 256x256. Image files used for training and testing are taken from http://openi.nlm.nih.gov/imgs/collections/ChinaSet_AllFiles.zip. Examples of some image samples used in this study can be seen in Figure 2 and Figure 3 below :

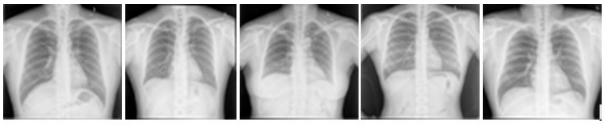

Figure 2.Examples of Negative Tuberculosis (TB) Image Samples

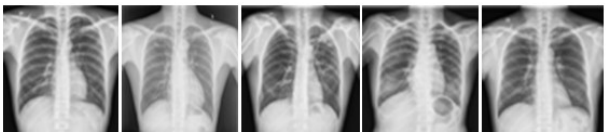

Figure 3.Examples of Positive Tuberculosis (TB) Image Samples

3.1 Image Extraction SURF

In the picture below, you can see an example of an x-ray image from the feature extraction of SURF:

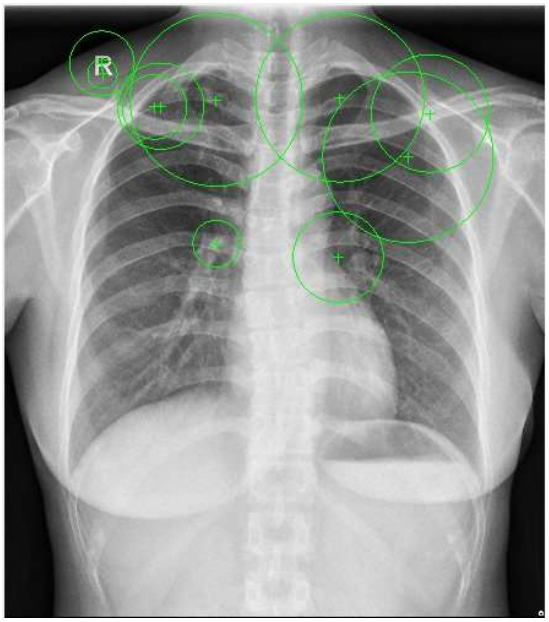

Figure 4. Image Extracted SURF

In this test, we present the results of the recognition of tuberculosis (TB) x-ray images using SURF feature

extraction and the K-Nearest Neighbor (KNN) classification method. The test results can be seen in Table 1 below:

X-ray Image Detection		
	Negative TB	***Positive TB***
Negative TB	69.7%	30.3%
Positive TB	23.33%	76.67%
Accuracy	**73.18%**	

Table 1 illustrates the results of the classification of negative and positive TB X-ray images using SURF feature extraction and the K-Nearest Neighbor (KNN) classification method, on testing negative TB x-ray images the accuracy of the K-Nearest Neighbor (KNN) method is 69.7% while On testing positive TB x-ray images, the accuracy of the K-Nearest Neighbor (KNN) method was 76.67%. The average accuracy of the K-Nearest Neighbor (KNN) method in this study is 73.18%.

The graph of the classification results of X-ray TB images using SURF feature extraction and the K-Nearest Neighbor (KNN) method in this study can be seen in Figure 5 below :

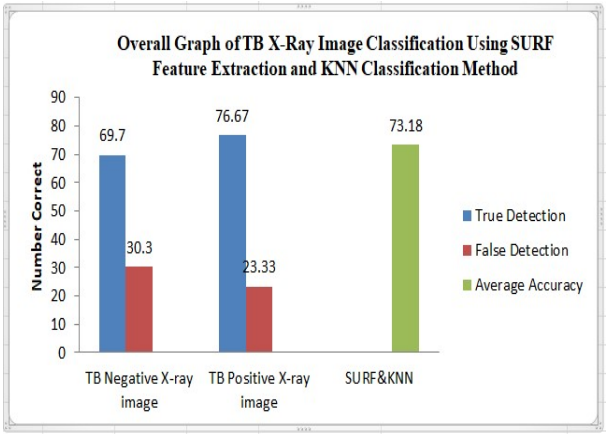

Figure 5.Graph of the overall classification of TB x-ray images using SURF feature extraction and the KNN classification method

4. Conclusion

Research [21] utilizes the KNN method and HOG feature extraction in classifying TB x-ray images with an average accuracy of 71.81%, whereas in this study using SURF feature extraction and the KNN classification method in classifying positive and negative tuberculosis (TB) the average accuracy of the KNN in this study was 73.18%. KNN with SURF extraction was 2% superior in the classification of TB x-ray images.

References

[1] Stacey Singer Leshinsky, "Pulmonary tuberculosis: Improving diagnosis and management," American Academy of Physician Assistants, vol. 29, no. 2, pp. 20-25, 2016.

[2] Christoph Lange et al., "Management of drug-resistant tuberculosis," vol. 394, no. 14, pp. 953-966, 2019.

[3] Akosua Adom Agyeman and Ofori-Asenso Richard, "Tuberculosis—an overview," Journal of Public Health and Emergency, vol. 1, no. 7, pp. 1-11, 2017.

[4] James J Dunn, Jeffrey R Starke, and Paula A Revella, "Laboratory Diagnosis of Mycobacterium tuberculosis Infection and Disease in Children," Journal of Clinical Microbiology, vol. 54, no. 6, pp. 1434-1441, 2016.

[5] Yeon Joo Jeong, Kyung Soo Lee, and Yim Jae-Joon, "The diagnosis of pulmonary tuberculosis: a Korean perspective," Precision and Future Medicine, vol. 1, no. 2, pp. 77-87, 2017.

[6] Aswindo Putra, Jondri Jondri, and Fitriyani Fitriyani, "Paralelisasi Klasifikasi Data Ekspresi Gen Kanker dengan Algoritma Deep Neural Network Menggunakan Stacked Sparse Autoencoder," e-Proceeding of Engineering, vol. 5, no. 2, pp. 8296-8310, 2018.

[7] Fitra Septia Nugraha, Muhammad Ja'far Shidiq, and Sri Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," Jurnal PILAR Nusa Mandiri, vol. 15, no. 2, pp. 149-156, 2019.

[8] Yennimar Yennimar and Reyhan Achmad Rizal, "Comparison of Machine Learning Classification Algorithms in Sentiment Analysis Product Review of North Padang Lawas Regency," SINKRON, vol. 4, no. 1, pp. 268-273, 2019.

[9] Reyhan Achmad Rizal and Christnatalis HS, "Analysis of Facial Image Extraction on Facial Recognition using Kohonen SOM for UNPRI SIAKAD Online User Authentication," SINKRON, vol. 4, no. 1, pp. 171-176, 2019.

[10] Indra Waspada, Adi Wibowo, and Noel Segura Meraz, "Supervised Machine Learning Model For Microrna Expression Data In Cancer," Journal of a Science and Information, vol. 10, no. 2, pp. 108-115, 2017.

[11] Chalifa Chazar and Bagus Erawan Widhiaputra, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma," Jurnal Informatika dan Sistem Informasi, vol. 12, no. 1, pp. 67-80, 2020.

[12] Agung Wibowo, "Aplikasi Diagnosis Penyakit Kanker Payudara Menggunakan Algoritma Sequential Minimal Optimization," Jurnal Teknologi dan Sistem Komputer, vol. 5, no. 4, pp. 153-158, 2017.

[13] Olha Musa and Alang Alang, "ANALISIS Penyakit Paru-Paru Menggunakan Algoritma K-

Nearest Neighbors Pada Rumah Sakit Aloei Saboe Kota Gorontalo," ILKOM Jurnal Ilmiah, vol. 9, no. 3, pp. 348-352, 2017.

[14] I Handayani, "Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column," Indonesian Journal of Information Systems, vol. 1, no. 2, pp. 57-66, 2019.

[15] M Syukri Mustafa and I Wayan Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi, vol. 7, no. 1, pp. 1-10, 2019.

[16] Yennimar Yennimar, Reyhan Acmad Rizal, Amir Mahmud Husein, and Mawaddah Harahap, "Sentiment analysis for opinion IESM product with recurrent neural network approach based on long short term memory," in International Conference of Computer Science and Information Technology (ICoSNIKOM), Medan, 2019.

[17] Arif Ridho Lubis, Muharman Lubis, and Al-Khowarizmi Al-Khowarizmi, "Optimization of distance formula in K-Nearest Neighbor method," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 1, pp. 326-338, 2020.

[18] Muhathir Muhathir, Reyhan Achmad Rizal, Julianus Stepanus Sihotang, and Rollys Gultom, "Comparison of SURF and HOG extraction in classifying the blood image of malaria parasites using SVM," in International Conference of Computer Science and Information Technology (ICoSNIKOM), Medan, 2019.

[19] Muhammad Baresi Ariel, Ratri Dwi Atmaja, and Azizah Azizah, "Implementasi Metode Speed Up Robust Feature dan Scale Invariant Feature Transform untuk Identifikasi Telapak Kaki Individu," Jurnal AL-Azhar Indonesia Seri Sains Dan Teknologi, vol. 3, no. 2, pp. 178-186, 2016.

[20] Mhd. Furqan Mhd. Furqan, Sriani Sriani, and Lailan Sofinah Harahap, "Klasifikasi Daun Bugenvil Menggunakan Gray Level Co-Occurrence Matrix dan K- Nearest Neighbor," Jurnal CoreIT, vol. 6, no. 1, pp. 22-29, 2020.

[21] Muhathir Muhathir, Theofil Tri Saputra Sibarani, and Al-Khowarizmi Al-Khowarizmi, "Analysis K-Nearest Neighbors (KNN) in Identifying Tuberculosis Disease (Tb) By Utilizing Hog Feature Extraction," vol. 1, no. 1, pp. 33-38, May 2020.