Analysis K-Nearest Neighbors (KNN) in Identifying Tuberculosis Disease (Tb) By Utilizing Hog Feature Extraction

Muhathir[1], Theofil Tri Saputra Sibarani[2], Al-Khowarizmi[3]
[1,2]Deparment of Informatics, Universitas Medan Area, Indonesia
[3]Department of Information Technology, Universitas Muhammadiyah Sumatera Utara, Indonesia

ABSTRACT

Pulmonary tuberculosis is an infectious disease caused by Microbacterium tuberculosis, which is one of the lower respiratory tract disease, which is largely in the pulmonary tissue of the lung infection and then undergoes a process known as the primary focus of Ghon. Because the disease is difficult and takes a long time to decide the patient is affected by the disease Tuberkolusis, then the detection of the patient affects Tuberkolusis by utilizing the K-NN method as a classification and HOG as feature extraction. Results of the classification of positive diagnosis with a total of 234 samples from 330 samples or successfully recognizable Sebasar 70.90%, while the classification result is a negative diagnosis with the amount of 240 samples from 330 samples or successfully identified by 72.72%. The results of the study showed the image classification of the X-ray Set Tuberculosis using the method K-NN and HOG feature with cross-validation 5 folds with 71.81% accuracy.

Keyword : tuberculosis, K-NN, HOG.

Corresponding Author:
Muhathir,
Department of Informatics, Faculty of Engineering
Universitas Medan Area,
Jalan Kolam No 1 Medan Estate / Jalan Gedung PBSI, Medan 20223, Indonesia.
Email: muhathir@staff.uma.ac.id

1. INTRODUCTION

Illness is an abnormal state of the body or mind that causes discomfort, tribulation against the person being influenced. One disease that is being a lot of discussion material is tuberculosis or commonly known as TB. The growing issue is the rate of patient growth faster than the number of doctors available. This is a big problem because every human being has the right to get good service for the illness that it suffered [1] the World Health Organization (WHO) stated that the current lung tuberculosis disease has become a global threat as nearly a third of the world's population has been infected[2].

Pulmonary Tuberitis is an infectious disease caused by the Microbacterium tuberculosis, which is one of the lower respiratory tract disease, which is largely in the pulmonary tissue of the lung infection, and subsequently undergo a process known as the primary focus of Ghon [3]. TB can attack anyone, especially the age of productive/still active work and children. Approximately 75% of TB patients are the most economically productive age group (15-50 years) [4]. The increasing number of tuberculosis sufferers is influenced by the number of poor people with unhealthy living patterns, unclean environment, and lack of information about the disease and its symptoms and causes that will make the treatment process slow. A slow and improper handling process will make the disease worse and fatal [5].

Related to the problem of disease management of TB, the task of a doctor will be very helpful when there is a system that can help the doctor in diagnosing TB disease. The purpose of the system is not to replace the role of a doctor, but rather to provide recommendations or possible diagnosis results based on the symptoms experienced by the patient. Also, the system is also built to help doctors to reduce the risk of human error due to the number of patients who have to be handled by a physician at one time [1].

In completing a conclusion, the diagnosis system can use certain methods to be implemented [1] As in the research conducted [6]. This study concluded the lung disease detection system designed in the study consisted of several parts of the system, namely pre-processing system, feature extraction

system and classification system. But the accuracy has not been satisfactory. Not much research on the predictions of TB disease. Subsequent studies [7] provide an accuracy value of 78.66% and an AUC value of 0.806 which identifies that the model is good classification.

Based on the above, it is necessary to develop a predictive system for the diagnosis of TB by using the KNN (K-Nearest Neighbors) method by utilizing HOG extraction, since the algorithm of K-Nearest Neighbor is easy to implement in diagnosing a disease, by utilizing HOG extraction, since the HOG method has been developed to detect other objects [8][9]. The algorithm K-Nearest Neighbor abbreviated KNN is usually applied in the classification of data based on the value of a small difference from the distance closest neighbor to the object. The general principle of this algorithm is to determine the value of K in the training data which will then be processed using KNN based on the distance. The next majority value of KNN is made basic in determining the class type or category of the next sample [10][11]. The extraction of a HOG feature which is one of the features of a panda image that has a good shape recognition capability [12], Histogram of Oriented Gradient (HOG) is the extraction of features used in an image processing computer by calculating the Gradient value on an image to get the result to be used to recognize the characteristics of that object [13].

So in this study, will discuss the performance of the K-NN algorithm in classifying the image of tuberculosis with the help of a Histogram of Oriented Gradient (HOG) feature extraction.

2. LETERATURE REVIEW

A. Tuberkulosis (Tbc)

Pulmonary Tuberitis is an infectious disease caused by Microbacterium tuberculosis, which is one of the lower respiratory tract disease, which is largely in the pulmonary tissue of the lung infection, and subsequently undergo a process known as the primary focus of Ghon [14]. Most TB germs invade the lungs, but it can also be related to other organs commonly referred to as extra lung TB. Lung TB is the most common form of about 80% of all sufferers. TB that attacks the lung tissue is the form of an easily contagious TB. TB extra lung is a form of TB disease that attacks the body organs other than the lungs. TB is essentially indiscriminately because this germ can attack all organs from the body [15]. People only know that TB is attacking the lungs only in general, but TB can also attack other organs besides the lung called extra lung. TB extra lung occurs when the TB germs spread to other organs of the body through the bloodstream. The definite diagnosis for TB disease is often difficult to be enforced while a working diagnosis can be enforced based on strong TB clinical symptoms (PRESUMTIF) by eliminating the possibility of other diseases [16].

B. Histogram of Oriented Gradients (Hog)

Histogram of Oriented Gradients (HOG) is a method used in image processing in order to detect objects [17]. The method was developed by Navneet Dalal and Bill Trigs in 2005 to detect pedestrians [18]. Histogram of Oriented Gradients (HOG) is a descriptor representing an object. The way the HOG works is by calculating the gradient value of a particular area of the image. Each image has the characteristic indicated by a gradient value obtained by dividing an image into the smallest area called cell [19].

According to [20] The Histogram of Oriented Gradient is the shape of the local object and the value used from the Gradient intensity. The process in using a HOG is to look for the gradient orientation and gradient vertical values and then look for the magnitude value and the cholinergic orientation of the original image size and then divide the image into several blocks that have a 2x2 size later in the block there are some cells with an 8x8 size that has the orientation of the gradient 9 bin so that it has a feature vector. To improve the performance of gradient values generated cholinergic orientation by normalizing in contrast, then this value is used to describe each block of the normalized value.

C. K-Nearest Neighbour (Knn)

The K-Nearest Neighbor (KNN) algorithm is a method of classifying the object based on the learning data that is closest to that object. Learning data is projected into multiple dimensioning spaces, each of which dimensions represent the features of the data [21][22][23].

The KNN algorithm includes methods that use the supervised algorithm [24][25][26]. The difference between supervised learning and unsupervised learning is that supervised learning aims to find new patterns in the data by connecting existing patterns of data with the new data. While on

unsupervised learning, data does not yet have any pattern, and the unsupervised learning objective is to find patterns in data. The goal of the KNN algorithm is to classify new objects by attribute and training samples [27][28]. Where the results of the test samples were newly classified based on the majority of categories on the KNN. In the classifying process, this algorithm does not use any model to match and is based solely on memory.

The goal of the KNN algorithm is to classify new objects based on the attributes and training samples. Clasifier does not use any model to match and is based solely on memory. Given a query point, it will be found a number of K objects or (training points) closest to the query point. Classification uses the most voting in the classification of K objects. The KNN algorithm uses a kinship classification as the predictive value of the new instance query. The KNN method algorithm is very simple, working based on the shortest distance from the instance query to the training sample to determine its KNN [29].

The best k value for this algorithm depends on the data. In general, a high k value will reduce the noise effect on the klsification, but makes the boundary between each classification increasingly blurred. A good k value can be chosen with the optimization of a parameter, for example by using cross-validation. A special case where the classification is reconditioned based on the closest training data (in other words, K = 1) is called the Nearest Neighbor algorithm.

3. RESEARCH METHOD

A. Dataset

The Data used in this weilitan is taken from Shenzhen Hospital X-Ray Set, The set contains images in JPEG format. There are 326 normal x-rays and 336 abnormal x-rays showing various manifestations of tuberculosis [30].

B. Research Steps

The research steps modeled in this study are illustrated in Figure 1.

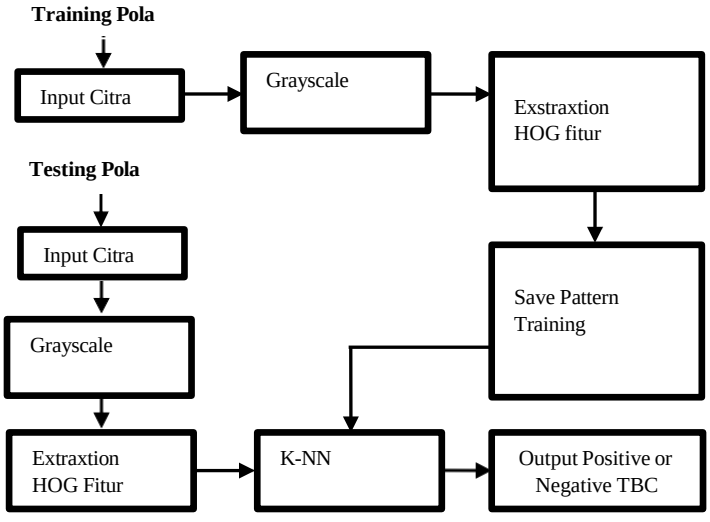


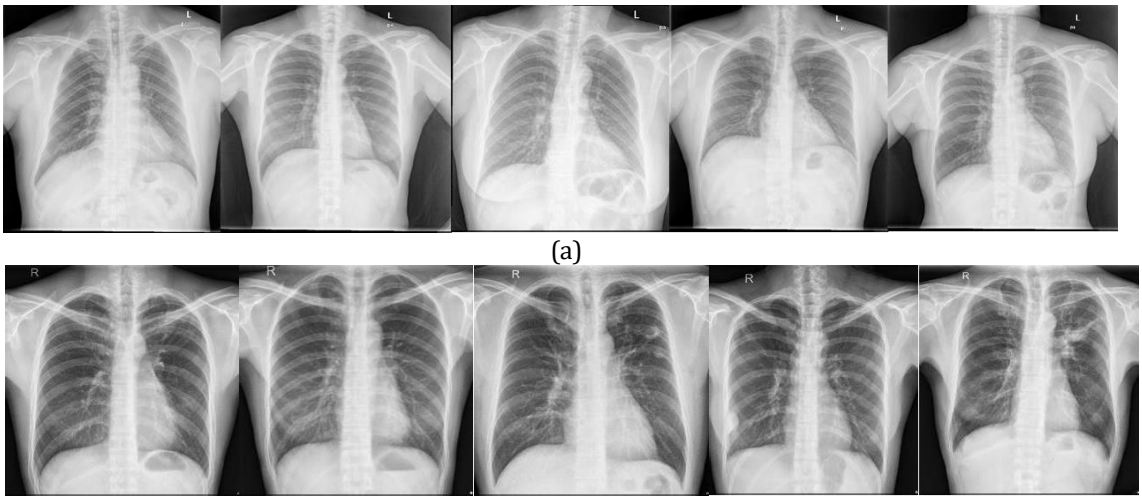
Fig 1. Generally Research Steps [31][32][33]

Figure 1 shows the research step that will be done by two processes, the first training process is the process of data processing (grayscale to minimize the colour space of the image of the three R,G,B color spaces into one color space i.e. grayscale as well as extracting by utilizing HOG features) as well as data stored as a pattern model to be used in the testing stage, both easy testing is the process of matching the model of the pattern that has been in training by utilizing the SVM method as a classification.

4. RESULTS AND DISCUSSION

A. Sampel X-ray Set Tuberculosis

This X-ray sample Set Tuberculosis was taken from the Shenzhen Hospital X-ray set, figure 2 attaching some samples to the X-ray Set Tuberculosis



(a)

(b)

Fig 2. Sampel X-ray Set Tuberculosis (a) Negative, (b) Positive.

B. HOG Detection

Hog feature detection results are marked with a cube in the image, Figure 3 shows the results of the Hog feature detection with 70 stronglest values in the Tuberculosis Set X-ray image..

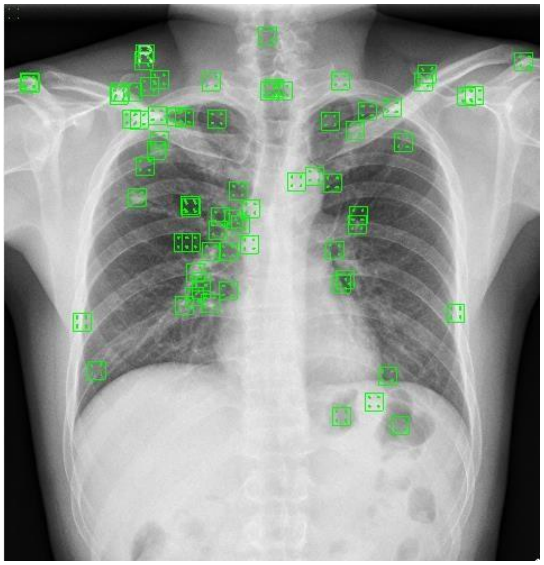


Fig 3. Deteksi HOG Fitur

C. Classification Result

Testing conducted in this study using cross-validation with 5 folds. Table 1 and Table 2 show the results of the X-ray Set Tuberculosis image classification by using the K-NN method with the help of HOG features.

Table 1. Classification results

	Diagnosis Positif	Diagnosis Negatif
Test Positif	234	96
Test Negatif	90	240

Table 2. Persentase classification results

	Diagnosis Positif	Diagnosis Negatif
Test Positif	70.90%	29.10%
Test Negatif	27.28%	72.72%

In table 1. Displaying TBC classification result by using K-NN and HOG feature extraction with sample amount, in a positive test of TBC classification result with a positive diagnosis with amount 234 and result of classification with negative diagnosis amounted to 96, while in a negative test of TBC classification with a negative diagnosis with number 240 and classification result with positive diagnosis amounted to 90. While in table 2. Displaying TBC classification result by using K-NN and HOG feature extraction by percentage, in a positive test of TBC classification result with a positive diagnosis with percentage 70.90% and the result of classification with negative diagnosis amounted to 29.10%, while on a negative test of classification with a negative diagnosis with percentage 72.72% and a result of classification with positive diagnosis amounted to 27.28%.

5. CONCLUSION

The results of the study showed the image classification of the X-ray Set Tuberculosis using the method K-NN and HOG feature with cross-validation 5 folds with 71.81% accuracy.

REFERENCES

[1] Wicaksono, B. S., Romadhony, A., & Sulistiyo, M. D. (2014). Analisis dan Implementasi Sistem Pendiagnosis Penyakit Tuberculosis Menggunakan Metode Case-Based Reasoning. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI).* Yogyakarta.

[2] Sarwani, D., Nurlaela, & Isnani, Z. A. (2012). Risk Factors of Multidrug Resistant Tuberculosis (MDR-TB). *Jurnal Kesehatan Masyarakat, 8*(1), 60-66.

[3] Muniroh, N., Aisah, S., & Mifbakhuddin. (2013). Faktor-Faktor Yang Berhubunga Dengan Kesembuhan Penyakit Tuberkulosis (TBC) Paru di Wilayah Kerja Puskesmas Mangkang Semarang Barat. *jurnal Keperawatan Komunitas*.

[4] Hiswani. (2008). *Tuberkulosis Merupakan Penyakit Infeksi yang Masih Menjadi Masalah Kesehatan Masyarakat.* Skripsi, Universitas Sumatera Utara.

[5] Shofia, E. N., Putri, R. R., & Arwan, A. (2017). Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.

[6] Fibrianto, A., Magdalena, R., & Fuadah, Y. N. (2018). KLASIFIKASI KONDISI PARU-PARU NORMAL, PENYAKIT TUBERKULOSIS (TBC) DAN EFUSI PLEURA PADA MANUSIA MENGGUNAKAN JARINGAN SYARAF TIRUAN PROPAGASI BALIK. *e-Proceeding of Engineering.*

[7] Wardani, R. S., & Purwanto. (2015). MODEL DIAGNOSIS TUBERKULOSIS MENGGUNAKAN k-NEAREST NEIGHBOR BERBASIS SELEKSI ATRIBUT. *The 2nd University Research Coloquium.*

[8] Nabilla, N. N., Hidayat, B., & Suhardjo. (n.d.). DETEKSI CITRA GRANULOMA MELALUI RADIOGRAF PERIAPIKAL DENGAN METODE HISTOGRAM OF ORIENTED GRADIENTS DAN KLASIFIKASI K-NEAREST NEIGHBOR. *Seminar Nasional Inovasi Dan Aplikasi Teknologi Di Industri* (p. 2018). Malang: ITN Malang.

[9] Prayudani, S., Hizriadi, A., Lase, Y. Y., & Fatmi, Y. (2019, November). Analysis Accuracy Of Forecasting Measurement Technique On Random K-Nearest Neighbor (RKNN) Using MAPE And MSE. In *Journal of Physics: Conference Series*(Vol. 1361, No. 1, p. 012089). IOP Publishing.

[10] Puspitawuri, A., Santoso, E., & Dewi, C. (2019). Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 3*(4).

[11] Lubis, A. R., Lubis, M., & Al-Khowarizmi (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326-338.

[12] Pranoto, M. B., & Ramadhani, K. N. (2017). Face Detection System Menggunakan Metode Histogram of Oriented Gradients (HOG) dan Support Vector Machine (SVM). *e-Proceeding Eng.*

[13] Cai, Z., Yu, P., Liang, Y., Lin, B., & Huang, H. (2016). SVM-KNN Algorithm for Image Classification Based on Enhanced HOG Feature. *Proceedings of the 4th IIAE International Conference on Intelligent Systems and Image Processing.*

[14] Shofia, E. N., Putri, R. R., & Arwan, A. (2017). Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.
[15] Aini, N., Ramadiani, & Hatta, H. R. (2017). SISTEM PAKAR PENDIAGNOSA PENYAKIT TUBERKULOSIS. *Jurnal Informatika Mulawarman, 12*(1).
[16] Alfaris, S. (2014). Sistem Pakar untuk Mendiagnosa Penyakit Polip Nasi (Polip Hidung) Menggunakan Metode Certainty Factor. *Pelita Informatika Budi Darma, 7*(2).
[17] Tanjung, J. P., & Muhathir. (2020). Classification of facial expressions using SVM and HOG. *JITE (JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING), 3*(2), 210-215.
[18] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, (pp. 886-893). San Diego, Calif, USA.
[19] SIDDIK, M. A., NOVAMIZANTI, L., & RAMATRYANA, I. N. (n.d.). Deteksi Level Kolesterol melalui Citra Mata Berbasis HOG dan ANN. *ELKOMIKA, 7*(2).
[20] Alamsyah, D. (2017). Pengenalan Mobil pada Citra Digital Menggunakan HOG-SVM. *Jatisi, 2*.
[21] Liantoni , F. (2015). Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *ULTIMATICS, 2*.
[22] Lubis, A. R., Lubis, M., Al-Khowarizmi & Listriani, D. (2019, August). Big Data Forecasting Applied Nearest Neighbor Method. In *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)* (pp. 116-120). IEEE.
[23] Al-Khowarizmi, Sitompul, O. S., Suherman & Nababan, E. B. (2017, December). Measuring the Accuracy of Simple Evolving Connectionist System with Varying Distance Formulas. In *Journal of Physics: Conference Series* (Vol. 930, No. 1, p. 012004). IOP Publishing.
[24] Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining.* New York: CRC Press.
[25] Larose, D. (2005). *Discovering Knowledge in Data.* USA: John Wiley's and Son.
[26] Han, J., & Kamber, M. (2006). *Data Mining:Concept and Techniques.* New York: Morgan Kaufmann Publisher.
[27] Sandi, B., Siahaan, J. K., Permana, P., & Muhathir. (2019). Klasifikasi Citra Wayang Dengan Menggunakan Metode k-NN & GLCM. *Semantika (Seminar Nasional Teknik Informatika). 2(1), 71-77.*
[28] Pariyandani, A., Larasati, D. A., Wanti, E. P., & Muhathir. (2019). Klasifikasi Citra Ikan Berformalin Menggunakan Metode k-NN dan GLCM. *Semantika (Seminar Nasional Teknik Informatika). 2(1), 42-47.*
[29] Al-Khowarizmi, A. K., Nasution, I. R., Lubis, M., & Lubis, A. R. (2020). The effect of a SECoS in crude palm oil forecasting to improve business intelligence. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1604-1611.
[30] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., & Thoma, G. (2014). *Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.* AME Publishing Company.
[31] Muhathir, Mawengkang, H., & Ramli, M. (2017). KOMBINASI Z-FISHER TRANSFORM DAN BRAY CURTIS DISTANCE UNTUK PENGENALAN POLA HURUF JAR PADA CITRA AL-QURAN. *Jurnal Bisman Info, 4*(1).
[32] Muhathir, M. (2018). KLASIFIKASI EKSPRESI WAJAH MENGGUNAKAN BAG OF VISUAL WORDS. *JITE (JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING), 1*(2), 73-82.
[33] Rizal, Fadlisyah, Muhathir, Akfal A.M. (2015). Detection System Tajwid Al Quran on Image Using Bray Curtis Distance. IJCAT, 2(8), 293-300.