Application of Data Mining for Tuberculosis Disease Classification Using K-Nearest Neighbor

Delima Sitanggang [1] [*], Lamria Simangunsong, [2] Geertruida Frederika Sundah [3], Rani Hutahaean [4], Indren[5]
Information Systems Study Program, Faculty of Science and Technology [1], [2], [3], [3],[4],[5]
Prima Indonesia University
Medan, Indonesia,

delimasitanggang@unprimdn.ac.id [1] lamriasimangunsong2@gmail.com[2],geertruidafrederika26@gmail.com [3]
ranihutahean059@gmil.com [4]indrenjaya2@gmail.com [5]

Abstrak - This study aims to determine how niali application of k-NN method and the value of accuracy obtained by k-NN method in clarifying the data of tuberculosis patients. This research focuses on improving public health and developing science to help people prevent and overcome tuberculosis. This type of research is qualitative. The study of used literature is the study of documentation. The algorithm uses the K-Nearest Neighbor algorithm. The results showed that the process of applying data mining to the classification of tuberculosis by using the K-Nearest Neighbor method got the final result of accuracy of 80%. Thus it can be concluded that the k-Nearest Neighbor algorithm is good.

Keywords— Tuberculosis, K-Nearest Neighbors, Classification

I. INTRODUCTION

One of the important vital organs in the human body is the lungs. The lungs are respiratory (breathing) organs that are affiliated with the respiratory system as well as flow (blood circulation) [1]. The main function of this organ is to exchange oxygen from the air with carbon dioxide from the blood. In the global world of health, lung disease is a challenge that cannot be ignored because it can result in death [2]. The triggering factors for lung disease are increasing air pollution, lifestyle changes, and other environmental factors [3]. Several types of diseases in the lungs are pneumonia, tuberculosis, bronchitis, and asthma. This study focuses on Tuberculosis disease in the lungs.

This disease has an impact on public health, because it is the main cause of death [4]. According to WHO data, the number of TB in Indonesia in 2020 is in third position with the burden of the highest number of cases[5]. Tuberculosis cases in Indonesia are estimated at 969,000 TB cases. In Indonesia, tuberculosis is a chronic disease that has become the No. 1 health problem in 2022, the Ministry of Health detected patients *Tuberculosis* (TB) more than 700,000 cases. In general, Tuberculosis is transmitted through the air that contains bacteria when TB patients actively cough and sputum, the bacteria will automatically be carried into the air and enter the body of other people who inhale [6]. There are many symptoms of tuberculosis, including chronic cough, fatigue, fever, to weight loss. Tuberculosis patients expel about 3000 sputum splashes when coughing and can last for hours in a dark or humid room. In most cases, people who breathe such air will develop Tuberculosis.

Tuberculosis is still a *problem* especially in the world of health, especially in countries that use high levels of poverty [7]. This disease affects all age groups with a very high incidence of the disease, an innovative approach is needed to analyze and handle this disease. Tuberculosis (TB) as a challenge focuses on the health sector of citizens around the world, including in Indonesia. Despite many prevention efforts, the spread of TB is still quite difficult to predict accurately. Therefore, a productive process is needed that can predict the spread of tuberculosis more accurately. One of the methods that can be used in this study is the K-NN (K-Nearest Neighbor) method.

The K-NN method is a simple algorithm used to classify data and regression in machine learning [8]. The K-NN method is one of the methods in the field of machine learning that can be used to classify data according to similarity using the nearest neighbor [9]. In the context of this study, K-NN will be applied to analyze and classify data of tuberculosis patients based on various clinical and laboratory parameters related to using this disease. Testing and validation of the K-NN model will be carried out using independent test data to evaluate the effectiveness and reliability of the method in classifying tuberculosis status in patients. This method is needed to make a significant contribution to improving the accuracy of assessment and helping early identification of tuberculosis problems, as a result of which it can support more effective prevention and treatment efforts [10].

Research GAP in this study refers to research [11] produce there was a significant relationship between contact with TB patients and the incidence of tuberculosis and OR.

II. LITERATURE REVIEW

A. Algoritma K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a machine learning algorithm that is non-parametric and lazy learning. A method that is non-parametric means that it does not make any assumptions about the distribution of the underlying data. In other words, there is no fixed number of parameters or parameter estimates in the model, regardless of whether the data is small or large.The K-Nearest Neighbor algorithm is used as

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/ sisfokom.v13i3.2218, Copyright ©2024
Posted : July 16, 2024, Revision : August 6, 2024, Received : August 14, 2024, Published : November 19, 2024

a dataset classification method based on previously classified training data. This involves supervised learning where the results of a new sample query are classified based on the proximity of the K-NN class [11]. KNN is the most basic and simplest classification technique, especially if there is little or no prior knowledge of the distribution of data. According to the KNN principle, each sample is classified similarly to the surrounding samples[12]. How close or distant the neighbors are usually calculated based on Eucledian distance. KNN is a widely used algorithm for text classification, relying on learning with training data to identify groups of objects [13].

K-Nearest Neighbor is one of the simplest algorithms used in machine learning for regression and classification. KNN follows a "bird of a feather" strategy in determining where new data should be placed. The KNN algorithm assumes that something similar will exist in close proximity or neighbors. This means that data that tends to be similar will be close to each other. KNN uses all available data and classifies new data or cases based on similarity size or distance function. The new data is then assigned to the class where most of the neighbor's data resides [14] .

B. Tuberculosis

Tuberculosis (TB) is a chronic disease that is transmitted through the air caused by bacteria *Mycobacterium* tuberculosis which can attack the organs of the body, especially in the lungs. Tuberculosis is an infectious disease caused by bacteria. There are 5 bacteria that are closely related to TB infection, namelyMycobacterium tuberculosis, Mycobacterium bovis, Mycobacterium africanum, Mycobacterium microti and Mycobacterium canettii [4].

III. RESEARCH METHODS

A. Collection of Data Sets

Information data on the types of diseases in the lungs consisting of 6 types of diseases consisting of Pneumonia, Bronchitis, Asthma, Other diseases in the lower respiratory tract, Pulmonary tuberculosis, Tuberculosis other than the lungs and other diseases in the upper respiratory tract but the researchers focused more on tuberculosis diseases obtained from the Darussalam Health Center. From 2019 to 2023, the author conducted research at the Darusallam Health Center and obtained the results of more than 1000 data.

B. Literature Studies

In literary research, theoretical sources related to research are collected in the form of magazines, books, articles and others, which support the research. Observation is an observation made to identify the material used for the research object, namely tuberculosis [15].

C. Data Acquistion

The TB dataset was obtained using primary data directly from the Darussalam Health Center. The data was processed through a data processing process and used to identify the cause of lung disease, namely tuberculosis.

D. Data Processign

The process of transforming raw data into an easy-to-understand and easy-to-understand form. This process is carried out because raw data often has an irregular format. The goal is to make it easier for the health center to see the patient data used by the researcher using data classification using the K-Nearest Neighbor Algorithm method as a source of information through the transferred collection to accurate and reliable data processing

IV. RESULTS AND DISCUSSION

A. Problem Analysis

TB is often difficult to diagnose because its symptoms are similar to those of other respiratory diseases, leading to delays in treatment. Long-term adherence to medication is necessary, but many patients stop treatment early, increasing the risk of drug resistance. Vaccination and health education programs are still ineffective, so technological approaches such as the K-Nearest Neighbor (K-NN) machine learning method are needed to analyze patient data, improve prediction accuracy, and support more effective TB prevention and treatment efforts.

B. Data Acquistion

The TB dataset was collected directly from the Darussalam Health Center. The data is processed through several stages such as collecting raw data from the source, changing the data format, filtering data to obtain relevant data.

Table 1. Data Acquisition

P	TBC	LUR	P-TBC	A	age	sex	YEAR
10	9	11	11	34	53	1	2019
30	18	23	26	17	37	1	2019
11	7	5	9	10	41	0	2019
13	8	12	9	10	52	1	2019
13	18	10	10	10	57	0	2019
13	8	12	30	20	57	1	2019
15	11	22	11	6	53	0	2019
36	10	8	13	13	44	1	2019
24	13	13	13	13	52	1	2019
24	13	11	13	13	57	1	2019
24	13	11	15	13	54	1	2019
13	13	17	36	20	48	0	2019
22	12	4	24	10	49	1	2019
9	12	4	24	10	64	1	2019
9	12	4	24	10	58	0	2019
9	12	4	13	23	50	0	2019
14	39	27	22	32	58	0	2019
16	2	3	9	13	66	0	2019
10	6	7	9	13	43	1	2019
10	6	7	9	1	69	0	2019
10	6	7	14	19	59	1	2019
13	11	15	16	18	44	1	2019
26	7	9	10	8	42	1	2019
17	13	8	10	8	61	1	2019
17	13	8	10	8	40	1	2019
17	13	8	13	27	71	0	2019
11	15	8	26	21	59	1	2019
9	39	8	17	7	51	1	2019
9	44	8	17	7	65	0	2019
11	54	27	17	7	53	1	2020

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/ sisfokom.v13i3.2218, Copyright ©2024
Posted : July 16, 2024, Revision : August 6, 2024, Received : August 14, 2024, Published : November 19, 2024

C. Uzi meaning

During the test, the first step was to process the TB data in Excel. After the data was processed and applied the K-Nearest Neighbor (K-NN) formula, the data was further processed using Matlab. The analysis included 143 data collected from 2019 to 2023.

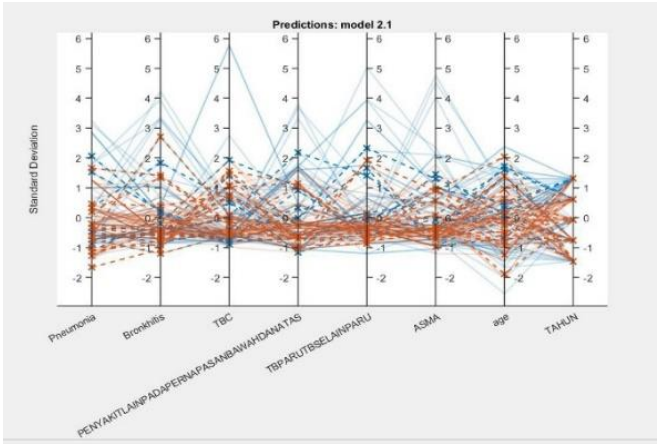


Figure 1. Disease Multidimensional Data Visualization

The figure above shows a visualization of the relationship between several variables in the disease dataset. The model used to predict this dataset has an accuracy of 80% and a total *misclasification cost* 30, meaning that the model can correctly classify data on 80% of all observations. The orange line color on the chart shows accurate accuracy while the blue color of the chart is the opposite of the orange color with inaccurate accuracy.

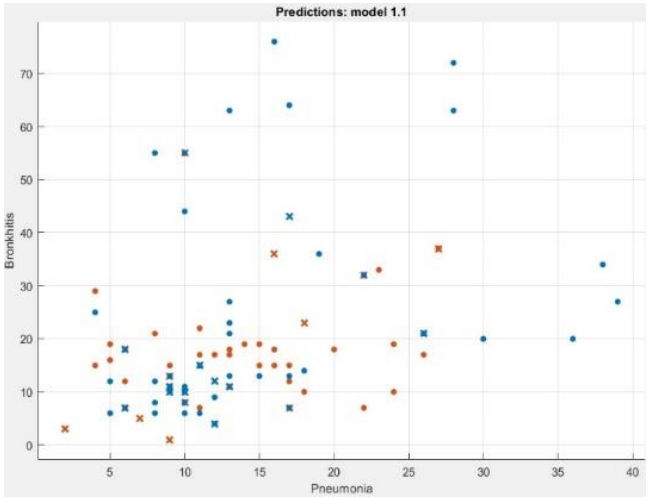


Figure 2. Matrix Evaluation

The figure above illustrates the relationship between two variables in the dataset. The model used to classify the relationship between these two variables has a data accuracy rate of 80%, which means that it is able to correctly classify data as much as 80% of all observations. The color of the dots on the plot image is used to indicate the predicted value of the model. These results show that the applied KNN model is effective in cassifying data with good accuracy, and additional metrics such as precision, Recal, and F2-Score provide deeper insights into the strengths and weaknesses of the prediction model.

Figure 3. Model Classification *Number of Observations*

Based on the figure, it can be seen that the model classification *number of observations* shows the number of true and false classifications of the model. In the scatter plot presented, a large number of data points are scattered following a linear pattern that shows a positive relationship between the X and Y variables, where the Y axis (*True Class*) and the X axis (*Predicted Class*) are known. On the Y axis 72 is true and 11 is false, while on the X axis 62 is true and 17 is false. Before modeling the training data in Matlab, the researcher first determined the average value of the dataset using excel to determine which one experienced the most tuberculosis in the past 5 years

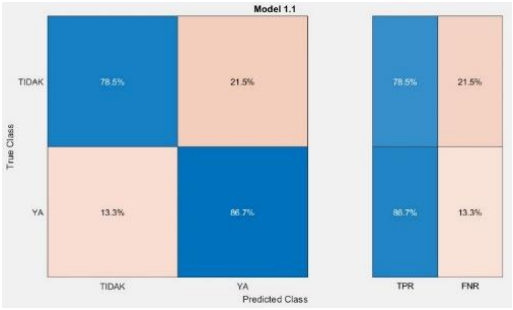
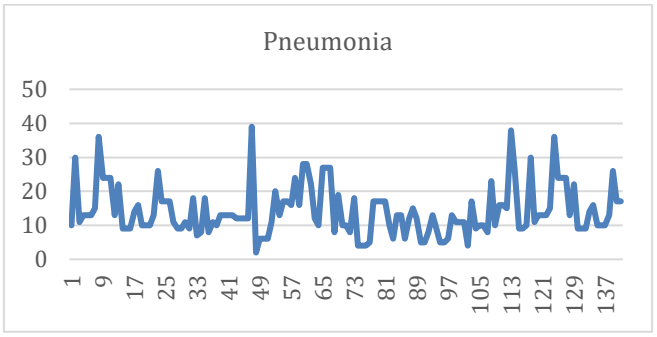


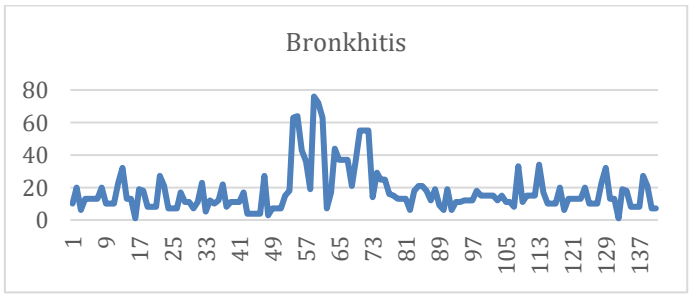
Figure 4. Model Classification *True Positive & False Negative*

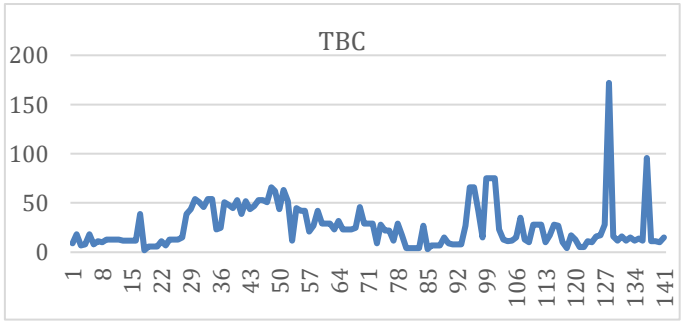
Based on the figure, the matlab classification modeling has the results of the overall TB disease data for the past 5 years with 6 attributes or types of TB disease with an accuracy result of 80%. The data is divided into training sets and test sets. And each type of disease data has different data accuracy. In the data, there are True Positive Rates or called TPR 78.5% and 86.7%, while for False Negative Rates or FNR are 21.5% and 13.3%. From the results of the above data, it can be seen that the data included in the K-Nearest Neighbor method by tuberculosis disease is more than the data that is categorized as

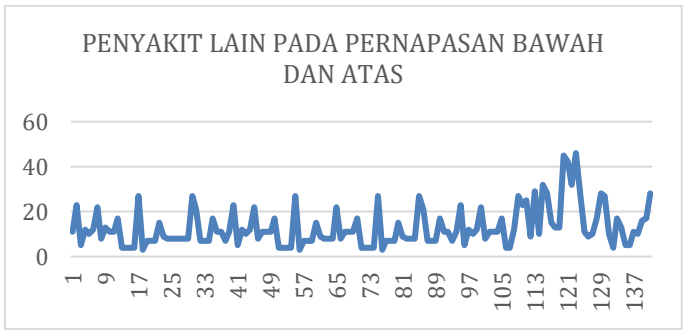
p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/ sisfokom.v13i3.2218, Copyright ©2024
Posted : July 16, 2024, Revision : August 6, 2024, Received : August 14, 2024, Published : November 19, 2024

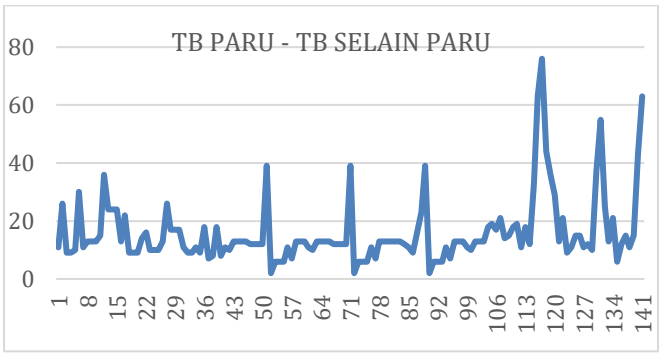
not included in the K-Nearest Neighbor method.











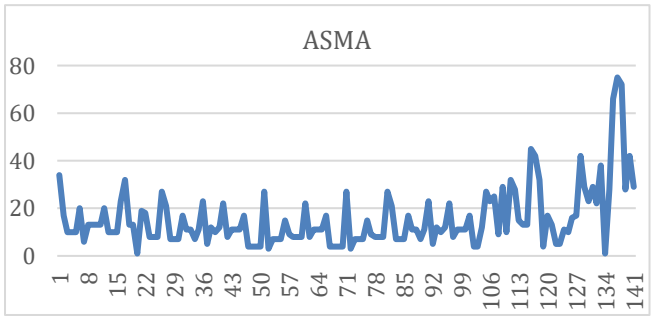


Figure 6. Data Visualization

V. CONCLUSION

The conclusion in this study refers to the process of applying data mining for the classification of tuberculosis disease using the K-Nearest Neighbor method to get a final result of 80% accuracy. Thus, it can be concluded that the K-Nearest Neighbor algorithm is good. In this study, the number of datasets used is still very small. Therefore, for the next research, it is necessary to add more datasets and use other algorithms such as Naive Bayes, Random Forest, Support Vector Machine, and others. By adding a larger dataset, it is hoped that the results of the study will be more accurate and representative. In addition, the application of additional algorithms will allow for a comparison of the performance of various methods, so that the most suitable algorithm for the case under study can be selected. Naive Bayes, Random Forest, and Support Vector Machine are some of the machine learning algorithms that have proven to be effective in various types of data analysis and classification, so the use of these algorithms can improve the quality and reliability of research results.

REFERENCE

[1] A. M. Argina, "Application of K-Nearest Neigbor Classification Method on Datasets of Diabetic Patients," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.

[2] E. T. Atok, D. R. Sina, and D. M. Sihotang, "Implementation of Case-Based Reasoning to Diagnose Tuberculosis Disease Using the K-Nearest Neighbor Algorithm," *J. Komput. and Inform.*, vol. 7, no. 2, pp. 124–128, 2019, doi: 10.35508/jicon.v7i2.1656.

[3] M. A. Brillianto, H. T. Fauzi, and T. S. Siadari, "Classification of tuberculosis and pneumonia in child x-ray images using statistical first order extraction method," vol. 8, no. 6, pp. 3271–3277, 2022.

[4] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Performance Analysis of Knn Method on Dataset of Patients with Breast Cancer," *Indones. J. Data*

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/ sisfokom.v13i3.2218, Copyright ©2024
Posted : July 16, 2024, Revision : August 6, 2024, Received : August 14, 2024, Published : November 19, 2024

Sci., vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[5] T. N. Halim, R. Martin, and ..., "Classification of Customer Satisfaction Towards E-Commerce Platforms with the K–Nearest Neighbor (K-NN) Method," *Jurassic (Jurnal Ris. ...*, vol. 8, pp. 512–523, 2023, [Online]. Available: http://ejurnal.tunasbangsa.ac.id/index.php/jurasik/article/view/636%0Ahttps://ejurnal.tunasbangsa.ac.id/index.php/jurasik/article/download/636/609

[6] Q. A. A'yuniyah and M. Reza, "Application of K-Nearest Neighbor Algorithm for Classification of Student Majors at SMA Negeri 15 Pekanbaru," *Indones. J. Inform. Res. Softw. Eng.*, vol. 3, no. 1, pp. 39–45, 2023, doi: 10.57152/ijirse.v3i1.484.

[7] H. Saleh, M. Faisal, and R. I. Musa, "Classification of Nutritional Status of Toddlers Using the K-Nearest Neighbor Method," *Simtek J. Sist. Inf. and Tek. Comput.*, vol. 4, no. 2, pp. 120–126, 2019, doi:10.51876/simtek.v4i2.60.

[8] I. S. Muallif, H. Budiman, and R. Natalis, "Application of Data Mining to Predict Stock Price Movements Using the K-Nearest Neighbor Algorithm," *J. Media Inform. Budidarma*, vol. 8, no. 1, pp. 497–507, 2024.

[9] L. Nur Aziza, R. Yuli Astuti, B. Akbar Maulana, and N. Hidayati, "Application of the K-Nearest Neighbor Algorithm for Food Security Classification in Central Java Province," *MALCOM Indones. J. Mach. Learn. Comput. Sci. J.*, vol. 4, no. 2, pp. 404–412, 2024.

[10] R. S. A, Y. Faturrahman, and A. Setiyono, "ANALYSIS OF RISK FACTORS FOR TUBERCULOSIS INCIDENCE IN THE WORKING AREA OF THE HEALTH CENTER OF NORTH CIPINANG BESAR VILLAGE, EAST JAKARTA ADMINISTRATIVE CITY," vol. 2, no. 4, pp. 346–354, 2021.

[11] H. A. Dwi Fasnuari, H. Yuana, and M. T. Chulkamdi, "Application of K-Nearest Neighbor Algorithm for Classification of Diabetes Mellitus," *Antivirus J. Ilm. Tech. Inform.*, vol. 16, no. 2, pp. 133–142, 2022, doi: 10.35457/antivirus.v16i2.2445.

[12] D. Sitanggang, N. Nicholas, V. Wilson, A. R. A. Sinaga, and A. D. Simanjuntak, "Implementation of Data Mining to Predict Heart Disease Using K-Nearest Neighbor and Logistic Regression Methods," *J. Tek. Inf. and Komput.*, vol. 5, no. 2, p. 493, 2022, doi: 10.37600/tekinkom.v5i2.698.

[13] W. Syahfira *et al.*, "Application of Case Based Reasoning Method for Disease Diagnosis in Cows," vol. 18, no. x, pp. 211–219, 1978.

[14] Y. Pratama, A. Prayitno, D. Azrian, N. Aini, Y. Rizki, and E. Rasywir, "Classification of Heart Failure Using the K-Nearest Neighbor Algorithm," *Bull. Comput. Sci. Res.*, vol. 3, no. 1, pp. 52–56, 2022, doi: 10.47065/bulletincsr.v3i1.203.

[15] I. D. Damayanti and A. Michael, "Determination of Coffee Fruit Maturity Level Using Image Histogram and K-Nearest Neighbor," *BAREKENG J. Science Mat. and Terap.*, vol. 18, no. 2, pp. 0785–0796, 2024, doi: 10.30598/barekengvol18iss2pp0785-0796.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/ sisfokom.v13i3.2218, Copyright ©2024
Posted : July 16, 2024, Revision : August 6, 2024, Received : August 14, 2024, Published : November 19, 2024