Scientific Journal of Informatics
Vol. 3, No. 2, November 2016
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

An Identification of Tuberculosis (TB) Disease in Humans using *Naïve Bayesian* Method

Agustin Trihartati S.[1], C. Kuntoro Adi[2]

[1,2]Department of Computer Science, Faculty of Science and Technology, Sanata Dharma University Yogyakarta
Email: [1]agustintrihartati@gmail.com,[2]kuntoroadi@usd.ac.id

Abstract

Tuberculosis (TB) is a disease that can cause a death if not recognized or not treated properly. To reduce the death rate of tuberculosis patients, the health experts need to diagnose that disease as early as possible. Based on the main indication data, laboratory test results and the rontgen photo, Naïve Bayesian approach in data mining techniques could be optimized to diagnose tuberculosis. Naïve Bayes classifiers predict class membership probabilities with a class that has the highest probability value. The output of the system is an identification Tuberculosis type of the patients. Testing of the system using 237 data sample with variation of cross-validation in 3, 5, 7 and 9-fold cross validation gives an average accuracy 85,95%.

Keyword: Naïve Bayesian, tuberculosis identification, cross-validation

1. INTRODUCTION

Tuberculosis is a direct infectious disease caused by TB bacteria (Mycobacterium tuberculosis). Most of the TB bacteria attack the lungs, but these can also attack other organs [1]. TB still becomes one of public health problems in the world. According to WHO report in 2013, it was estimated there were 8.6 million cases of TB in 2012 in which 1.1 million the dead (13%) were TB patients with HIV-positive. Approximately 75% of the patients were in the African region. Whereas, child mortality (with HIV-negative status) suffered from TB reached 74,000 deaths / year, or for about 8% of the total deaths caused by TB [2].

TB can be classified by anatomic location or locations of organs afflicted by TB [2], namely: (1) Pulmonary Tuberculosis, it is TB that occurs in *parenchyma* (lung tissue). Miliary TB is considered as pulmonary TB because of lesions in the lung tissue. Chest cavity TB lymphadenitis (hilum or mediastinum) or pleural effusion without radiologist depiction that supports TB in the lungs are stated as extra-pulmonary TB. Patients suffering from pulmonary tuberculosis and suffering from extra-pulmonary TB as well are classified as pulmonary TB patients. (2) Extra-Pulmonary Tuberculosis, this is TB that occurs in organs other than the lungs, e.g.: pleura, lymph node, abdomen, urinary tract, skin, joint, brain membrane and spine. The diagnosis of extra-pulmonary TB can be determined by the result of bacteriological or clinical checkups. Extra-Pulmonary TB diagnosis should be pursued based on the discovery of Mycobacterium tuberculosis. Extra-Pulmonary TB patients with TB at some organs are classified as Extra-Pulmonary TB patients.

TB generally occurs in the lungs (pulmonary TB). However, the spread through the bloodstream or lymph can cause TB other than the lungs organs (Extra-Pulmonary

TB).The main symptom of TB patient is coughing up phlegm for two weeks or more. Cough may be followed by additional symptoms namely sputum mixed with blood, coughing up blood, shortness of breath, weakness, decreased appetite, weight loss, malaise, sweating in the night without physical activity, and fever more than a month.

As the number of TB patients is still high, everybody who comes up with symptoms above is regarded as an unexpected TB patient. Hence, it is essential to conduct microscopic examination of sputum. The aim of sputum examination is to establish the diagnosis. If the microscopic examination is negative, then an examination of X-rays that supports the existence of tuberculosis should be done [2].

In 2013, there was a suggestion of some countries of WHO members which proposed a new strategy for controlling TB that was able to withstand the rate of new infections and prevent deaths from TB. The strategy was outlined in the three main pillars of the strategy and its components. One of the components was the TB diagnosis as early as possible [2]. This research identified TB disease as early as possible through the patients' primary symptoms, the results of sputum examination in the laboratory and X-rays based on the book of the National Guidelines for Tuberculosis Control (2014). As the grouping result, the research classified some groups based on anatomic location or locations of organs afflicted by TB namely Pulmonary TB, Extra-Pulmonary TB and non TB.A *Naïve Bayesian data mining* technology was used to process the data of symptoms, the sputum examination results in the laboratory and X-rays results. *Naïve Bayesian* method will calculated the probability on each case value of the target attribute in every sample data. Furthermore, *Naïve Bayesian* classified the sample data to a class that had the highest probability value. The implementation of Naïve Bayesian method was expected to help calculate probabilities on the sample data to identify tuberculosis (TB) disease in humans.

2. METHOD

2.1 *Naïve Bayesian*

Naïve Bayesian classification is one of the algorithms contained in classification techniques. It is a method based on bayes theorem found by Thomas Bayes.Itpredicts future opportunities toward previous experiences by using probability and statistics methods. The equation of Bayes' theorem is [3]:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

in this case:

X	= data with unknown class (the set of training data)
Y	= hypothesis
(Y|X)	= posterior probability, that is conditional probability of hypothesis Y based on condition X.
(Y)	= prior probability of hypothesis Y, namely probability that the hypothesis Y is valid before data X appears.

(X) = probability of data X
(X|Y) = conditional probability of X based on the condition of hypothesis Y, and it is called likelihood.

Naïve Bayesian Classification assumes that the influence of attribute value in a certain class does not depend on the values of other attributes. The condition is expressed by the formula as follows [4]:

$$P(Y|X) = \frac{P(X1 \mid Y)P(X2 \mid Y)..P(Xn \mid Y)P(Y)}{P(X)} \quad (2)$$

Description:

X	=	the set of training data
Y	=	hypothesis
(Y|X)	=	prior probability of hypothesis Y, namely conditional probability of hypothesis Y based on condition X
(Y)	=	prior probability of hypothesis Y, namely probability in which hypothesis Y is valid before data X appears
(X)	=	probability of data X.

$P(X1 \mid Y)P(X2 \mid Y)..P(Xn \mid Y)P(Y)$ = Probability of X1, X2, Xn for hypothesis Y, it is usually called likelihood.
Since P(X) is irrelevant, hence, it only uses the following formula to seek likelihood [4]:

$$P(Y|X) = P(X1|Y)P(X2|Y)..P(Xn|Y)P(Y) \quad (3)$$

If the value of P(Xn|Y) is 0, so P(Y|X) value = 0. Thus, Naïve Bayesian classification cannot be done, because Naïve Bayesian classification cannot predict the record in which one of its attributes has conditional probability (likelihood) = 0. To overcome that problem, it is done adding a value of 1 to any evidence / P (X) in the calculation so that the probability will not be 0 [5].

2.2 *Cross Validation*

In the cross validation approach, each of the data is used in the same number of training and is used once for testing. The general form of this approach is called the *k-fold cross validation*, which breaks a set of data into k parts of data sets in the same size. Each time test conducted, the fraction is used as a set of test data while the other fraction is used as a set of training data. The procedure is done as much as k times so that each data has a right test data opportunity and becomes a training data as k-1 times. The total error is obtained by summing all errors attained from k-time process [6].

2.3 Accuracy using Matrixs Confusion

Matrixs confusion is a table that records the classification result. The example of Matrixs confusion is shown in the Table 1:

Table 1. Matrixs Confusion Example

f_{ij}		Predicted result class (j)	
		class = 1	class = 0
Original class(i)	class = 1	f_{11}	f_{10}
	class = 0	f_{01}	f_{00}

Table 1 is an example of Matrixs confusion that classifies binary (two-class) for two classes, such as class 0 and 1. Each fij cell in matrix states the number of records / data from class i that its predictions come in class j. For example f11 cell is the number of data in class 1 which is properly mapped to class 1. F10 cell is the data in class 1 which erroneously mapped to class 0.Based on Matrixs confusion content, it can be seen that the number of properly predicted data in each class is (f11 + f00) and erroneously classified data is (f10 + F01). Matrixs confusion quantity can be summarized into two values, namely accuracy and error rate. By knowing the number of properly classified data, the accuracy of predicted result is detected. Besides, by knowing the number of erroneously classified data, the error rate from the prediction can be detected. These two quantities are used as classification performance matrix. To calculate the accuracy, the following formula is used [6]:

$$accuracy = \frac{\text{number of properly predicted data}}{\text{number of conducted prediction}} = \frac{f11+f00}{f11+f10+f01+f00} \quad (4)$$

2.4 Data

The data used in this study came from several health centers in Kulon Progo, Yogyakarta. It was taken from the medical records of patients who were diagnosed pulmonary TB, Extra-Pulmonary TB, and non TB. It had nine attributes that contain of symptoms data (cough for more than 2 weeks, asthma, fever, weight loss, coughing up blood). The laboratory results data labeled A, B, C, and the last attribute was the X-ray result. Besides, there was one class that contained pulmonary TB, Extra-Pulmonary TB, and non TB. Research was done by different feature / attributes. The Features were symptom, lab, symptom and lab, symptom and X-ray, lab and X-ray, symptom, lab and X-ray.

2.5 Research Process

Research diagram block can be seen in Figure 1.

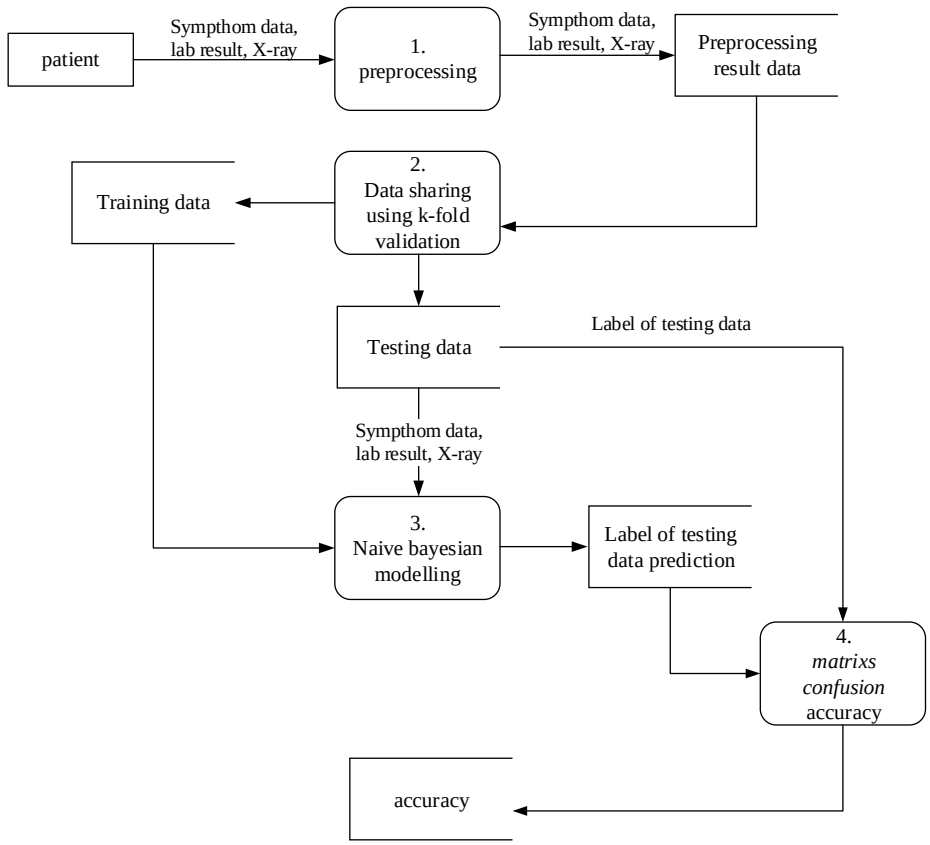

Figure1. Diagram block of TB research identification

Here is an explanation of each step in the process of test instrument.

1) The preprocessing stage did some of the following: data cleansing, data integration, data selection, and data transformation. The data in the string format was converted into data in the form of numbers or labels.
2) After passing through the preprocessing stage, the data was divided into training data and testing data. The data division used cross validation which was k-fold validation. The tests used 4 kinds of k-fold, namely 3, 5, 7, and 9 fold. Each test took 1 testing data and the other data was used as training data. In the next tests, the testing data was the training data and vice versa. Training data and training label were used to build the model, while testing data was used to test the model.
3) In modeling stage, the probabilities calculation for each training data was conducted based on training label. The formula used to calculate the probability is as follows:

$$P(Y|X) = P(X1|Y)P(X2|Y)..P(Xn|Y)P(Y) \quad (5)$$

It calculates the hypothesis probability Y based on condition X. Identification result derives from the calculation of the highest probability.

4) After passing calculation process by using Naïve Bayesian, the accuracy of the system was tested by using testing data. The calculation of accuracy used Matrixs Confusion, by summing the correct data and dividing it with either correct or wrong data and multiplying 100%. Table 2 is Matrixs Confusion table.

Table 2. Matrixs Confusion Table

Class	Pulmonary TB	Extra-Pulmonary TB	Non TB
Pulmonary TB	A	B	C
Extra- Pulmonary TB	D	E	F
Non TB	G	H	I

Based on Table 2, the number of the accuracy of each test was calculated using the following accuracy formula, as followed:

$$accuracy = \frac{\text{number of properly predicted data}}{\text{number of predictions made}} \quad (6)$$

The number of properly predicted data was located on a diagonal line that was data A + E + I. The number of prediction made was the sum up of all matrix confusion calculation result namely A + B + C + D + E + F + G + H + I. After that, the accuracy value was multiplied by 100%.

The test of data testing was done based on the number k used. Here are testing flowcharts diagram.

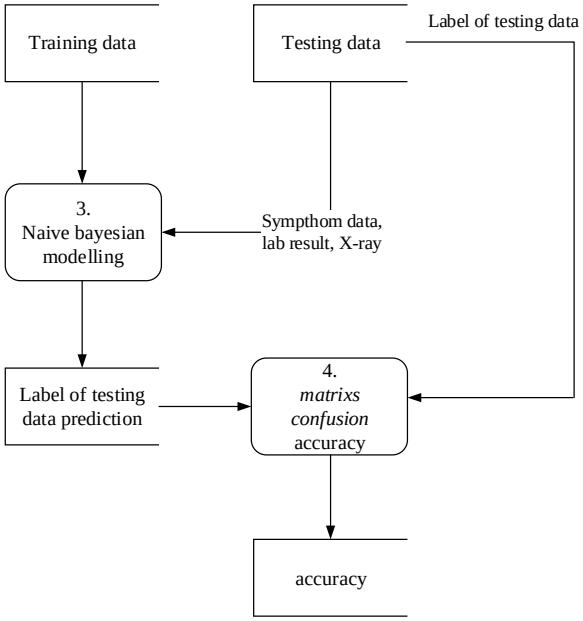

Figure2. Testing flowchart diagram

Every testing data passed the testing process in accordance with the testing flowchart diagram. Here is testing table conducted in accordance with the number of fold used.

Table3. The example of 3 fold division

Examination	Training	Testing
1	1, 2	3
2	1, 3	2
3	2, 3	1

3. RESULTS AND DISCUSSION

Based on the results of tests using Naïve Bayesian classification and k-fold validation, it was obtained the accuracy results as shown in Table 4.

Table 4. Test result table

Feature/Method	Accuracy %	Average Accuracy %
Symptom		
3 Fold	58,29	58,08
5 Fold	57,44	
7 Fold	57,45	
9 Fold	59,14	
Lab		
3 Fold	68,08	68,08
5 Fold	68,08	
7 Fold	68,08	
9 Fold	68,08	
Symptom and Lab		
3 Fold	70,63	69,3575
5 Fold	68,08	
7 Fold	69,36	
9 Fold	69,36	
Symptom and X-ray		
3 Fold	69,36	68,6125
5 Fold	68,08	
7 Fold	68,08	
9 Fold	68,93	
Lab and X-ray		
3 Fold	85,95	85,95
5 Fold	85,95	
7 Fold	85,95	
9 Fold	85,95	
Symptom, Lab and X-ray		
3 Fold	85,53	85,21
5 Fold	85,53	
7 Fold	84,68	
9 Fold	85,1	

From the twenty-four tests with different data usage and the number of different folds, it was obtained different classification and accuracy results. The highest accuracy result was 85.95% by using a feature lab and X-ray for 3 fold, 5 fold, 7 fold, and 9 fold which produced the same accuracy. The lowest accuracy result was 57.44 by using feature symptom and using 5 fold. The average result graph of accuracy test is shown in Figure 3.

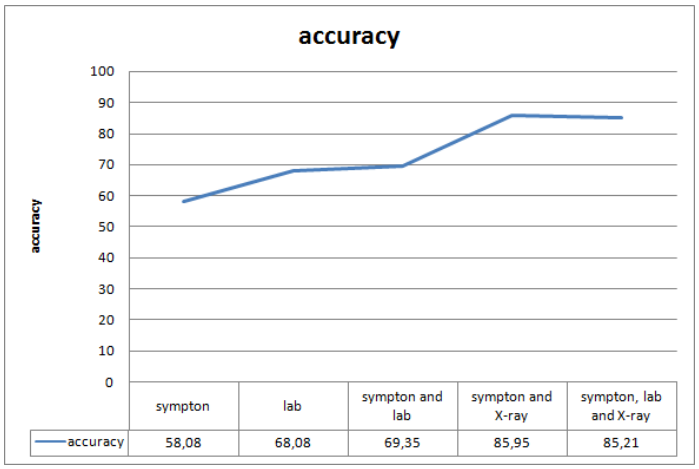

Figure3. Average accuracy result graph

From Figure 3, it was shown that the highest accuracy of this study was the test using the feature / attribute lab and X-ray. While the lowest acurracy average result was the test using the feature / attribute symptom. Feature symptom usage gave the lowest accuracy average result. It was bacause the main symptoms of TB patients such as cough for more than 2 weeks, asthma, weight loss, fever, coughing up blood were also found in lung diseases other than TB. The examples are bronchiectasis, chronic bronchitis, asthma, lung cancer, and etcetera. Therefore, it needed more supporting data such as the laboratory inquest result of phlegm and X-ray result. These data greatly affected the classification process. It was seen from the average accuracy result using feature/attribute lab and X-ray that produced 85.95% which was the highest accuracy in this study.

4. CONCLUSION

From these results, it is concluded that the Naïve Bayesian method can automatically identify tuberculosis well. The feature / attribute selection used for classification process affects the accuracy of the classification. From the 24 tests, it is obtained the highest average accuracy result 85.95% by using lab result feature and X-ray. The lowest average is 58.08 by using symptom feature.

5. REFERENCES

[1] DinasKesehatan Daerah Istimewa Yogyakarta. (2015). *Petunjuk Teknis Manajemen TB Anak*. Jakarta : Kementrian Kesehatan RI.

[2] Kementrian RI Direktorat Jenderal Pengendalian Penyakit dan Penyehatan Lingkungan. (2014). *Pedoman Nasional Pengendalian Tuberkulosis*. Jakarta: Kementrian Kesehatan RI.

[3] Tan, P. N., Steinbach. M., Kumar, V. (2006).*Data Mining: Introduction To Data Mining*. Boston: Pearson Addison Wesley.

[4] Han, Jiawie., Kamber, M. (2006). *Data Mining: Concepts and Technique* Second Edition. Morgan Kaufman Publishers, Amsterdam.

[5] Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta : GrahaIlmu.
[6] Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.