SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Diagnosis of Tuberculosis by Artificial Neural Network Algorithm

Amrin
Universitas Bina Sarana Informatika
Fakultas Teknologi Informasi
Jakarta, Indonesia
amrin.ain@bsi.ac.id

Abstract— Tuberculosis is an infectious disease caused by a bacterium called Mycobacterium tuberculosis and is the highest cause of death that occurs in productive age 15-50 years, weak economic groups, and low educated. In this study, the author will apply the data mining classification method, namely the Artificial Neural Network Algorithm to diagnose tuberculosis. Based on the results of the performance measurement of the model using Cross Validation, Confusion Matrix and ROC Curve testing methods, it is known that artificial neural network algorithms have an accuracy rate of 89.89% and area under the curva (AUC) value of 0.975. This shows that the resulting model including the classification category is very good because it has an AUC value between 0.90-1.00.

Keywords: Artificial Neural Network, confusion matrix, ROC curve

I. INTRODUCTION

Tuberculosis is an infectious disease caused by dropet nuclei when a tuberculosis patient coughs and sprinkles of saliva containing bacteria are inhaled by others while breathing [17]. Tuberculosis [12] is an infectious disease caused by a bacterium called Mycobacterium tuberculosis and is the highest cause of death that occurs in productive age 15-50 years, weak economic groups, and low educated. According to the Indonesian Ministry of Health in [3] tuberculosis can be contagious so it needs intensive treatment, at least 6 months of routine and continuous treatment is needed. While Indonesia is ranked second in the world after India with the most tuberculosis patients and it is estimated that there are 1.020,000 cases of tuberculosis in Indonesia.

According to the Indonesian Ministry of Health in [3] Transmission of tuberculosis is very rapid through the air. For sufferers, they are expected to always carry out examination and treatment completely. Tuberculosis is transmitted by air. The sprinkling of spit or phlegm that is released becomes a very fast transmission media in this world. Transmission of tuberculosis through the air will be very vulnerable to occur in public spaces. From various studies there will be tens of thousands of germs that come out of coughing and sneezing. Therefore it is hoped that the community will use masks in public places and always have a clean and healthy lifestyle.

Classification of tuberculosis disease data on medical is an important task in predicting disease, it can even help doctors in making decisions about the diagnosis of the disease [4], so it is very important to diagnose early in order to reduce transmission of tuberculosis to the general public. In this study, the author will apply the data mining classification method, namely the Artificial Neural Network Algorithm to diagnose Tuberculosis. The data that the authors use is data on patients at the Puskesmas Bojonggede diagnosed with tuberculosis.

II. LITERATURE REVIEW

A. Data Mining

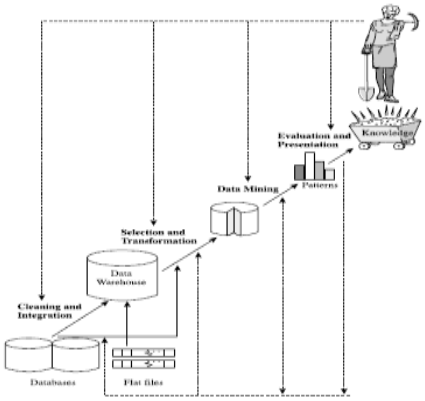
According to Han and Kamber in [2] Data mining is a series of processes to explore added value in the form of unexplored information from a database, explore with certain ways to manipulate data into more valuable information by extracting and recognizing important patterns from database. According to Daryl Pregibons in [5] "Data mining is a blend of statistics, artificial intelligence, and database research". The name data mining comes from the similarity between searching valuable information from a large database and mining a mountain for something of value [15]. Both require

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

screening through a large amount of material, or investigating intelligently to find the existence of something called value earlier.

Data Mining is a new technology that is very useful to help companies find very important information from their data warehouses. Some data mining applications focus on predictions, they predict what will happen in new situations from data that describes what happened in the past [18].

Data mining is often called Knowledge Discovery in Databases or abbreviated as KDD, is an activity that includes the collection, use of historical data to find order, pattern or relationship in large data sets [13]. The picture of the stages of making a data mining application is shown in figure 1 below:



Source: Han & Kamber

Figure 1. Steps in the KDD Process

Figure 1 shows the steps in the data mining process. The process in the data mining stage consists of three main steps, namely [14]:

1. Data Preparation

In this step, the data is selected, cleaned, and done preprocessed following the guidelines and knowledge of domain experts who capture and integrate internal and external data into the organization's overall review.

2. Data mining algorithm

The use of data mining algorithms is done in this step to explore integrated data to facilitate identification of valuable information

3. Data analysis phase

Output from data mining is evaluated to see whether the knowledge domain is found in the form of a rule that has been extracted from the network.

B. Artificial Neural Network

Artificial neural network is one of the artificial representations and human brain that always tries to simulate the learning process in the human brain. The term artificial here is used because this neural network is implemented using a computer program that is able to complete a number of calculation processes during the learning process [7]. The neural network is a set of connected input / output units where each relation has a weight.

Neural Networks are intended to simulate the behavior of biological systems of human nervous systems, which consist of a large number of processing units called neurons, which operate in parallel [1]. Neurons have a relationship with synapse that surrounds other neurons. The nervous system is presented in a neural network in the form of a graph consisting of nodes (neurons) that are connected by an arc, which corresponds to synapse. Since the 1950, neural networks have been used for prediction purposes, not only classifications but also for regression with continuous target attributes [16].

Neural networks consist of two or more layers, although most networks consist of three layers: input layer, hidden layer, and output layer [8]. The neural network approach is motivated by biological neural networks. Roughly speaking, the neural network is a set of connected input / output units, where each connection has the weight associated with it. Neural networks have several features that make them popular for clustering. First, neural networks are inherently parallel and distributed processing architectures. Second, learning neural networks by adjusting the interconnection weights with data, this allows the neural network to "normalize" patterns and act as extractors' features for different groups. Third, neural networks process numerical vectors and require object patterns to be represented by quantitative features only [5].

Multi Layer Perceptron also called multilayer feedforward neural network is a class of neural network Backpropagation algorithm for multilayer perceptron, is a systematic method for training so that it can be done and more efficient [10]. MLP consists of input layers, one or more hidden layers, and an output layer. The learning step in the backpropagation algorithm is as follows [11]:

1. Initialize network weights randomly (usually between -0.1 to 1.0)

2. For each data in the training data, calculate the input for the node based on the current input value and network weight, using the formula:

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$Input_j = \sum_{i=1}^{n} O_i w_{ij} + \square_j$$

Information:
Oi = Output node i from the previous layer
wij = the weight of the relation from the i node in the previous layer to node j
θj = error (as a delimiter)

3. Based on the input from step two, then generate the output for the node using the sigmoid activation function:

$$Output = \frac{1}{1 + e^{-Input}}$$

4. Calculate the Error value between the predicted value and the actual value using the formula:

$Error_j = Output_j \cdot (1 - Output_j) \cdot (Target_j - Output_j)$

Information:
Outputj = Actual output from node j
Target = Known target value on training data

5. After the Error value is calculated, then it is reversed to the previous layer (backpropagated). To calculate the value of an error in the hidden layer, use the formula:

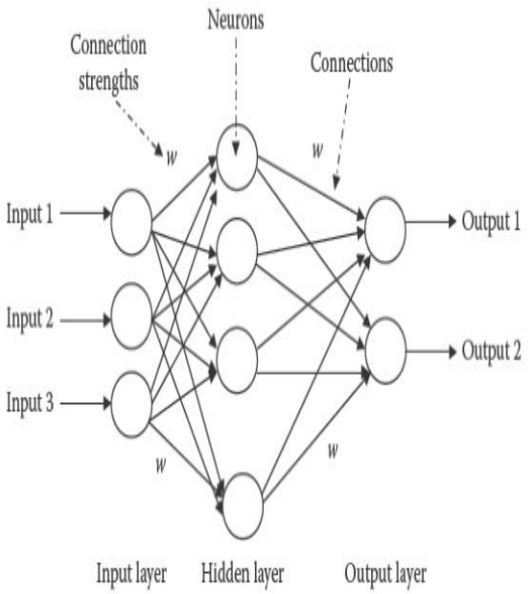
$$Error_j = Output_j(1 - Output_j) \sum_{k=1}^{n} Error_k \, w_{jk}$$

Information:
Outputj = Actual output from node j
Errork = error node k
wjk = Weight of relation from node j to node k in the next layer

6. The value of the error generated from the previous step is used to update the weight of the relation using the formula

$w_{ij} = w_{ij} + l \cdot Error_j \cdot Output_i$

Information:
wij = the weight of the relation from unit i in the previous layer to unit j
l = learning rate (constant, the value is between 0 and 1)
Errorj = Error at node output layer j
Outputi = Output from node i



Source: Shukla et al.
Figure 2: Architecture of the Neural Network

C. Evaluation and Validation of the Model

To measure the accuracy of the model evaluation and validation is carried out using techniques:

1. Confusion matrix

Confusion Matrix is a visualization tool commonly used in supervised learning. Each column in the matrix is an example of a prediction class, while each row represents an actual class event [5]. The confusion matrix contains actual and predicted predicted information on the classification system.

2. ROC (Receiver Operating Characteristic) Curve

The ROC curve shows accuracy and compares classification visually. ROC expresses confusion matrix. ROCs are two-dimensional graphs with false positives as horizontal lines and true positives as vertical lines [16]. The area under curve (AUC) is calculated to measure the difference in performance of the method used. AUC is calculated using the formula: [9]

$$\theta^r = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi\,(\mathrm{x}i^r, \mathrm{x}j^r)$$

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Where

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

Performance of AUC accuracy can be classified into five groups, namely [5]:
0.90 - 1.00 = Exellent Clasification
0.80 - 0.90 = Good Clasification
0.70 - 0.80 = Fair Clasification
0.60 - 0.70 = Poor Clasification
0.50– 0.60 = Failure

III. RESEARCH METHOD

This study consists of several stages as seen in the framework of thinking Figure 3 Problems (problems) in this study are not yet known accurate algorithms for diagnosing tuberculosis.

For this reason artificial neural network approach (model) was made to solve the problem and then tested the performance of the three methods. Testing uses the Cross Validation method, Confusion Matrix and ROC curve. To develop applications (development) based on the model created, Rapid Miner is used.

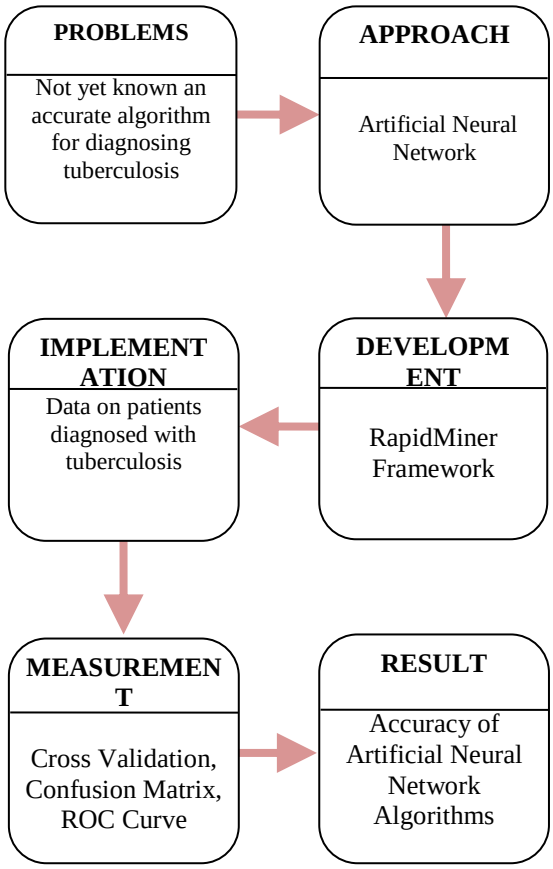


Figure 3. Framework for Problem Solving

IV. DISCUSSION

A. Data Analysis

In this study the data used were 136 tuberculosis patient data both positive and negative. Input variables in this study consisted of six variables, namely: 1. Sweat at night without physical activity, 2. Weight loss, 3. Decreased appetite 4. Easy to get tired and weak, 5. Fever, 6. Cough with phlegm more than three weeks accompanied by coughing up blood, while the output variable is variable TBC disease. The software used to analyze is RapidMiner version 5.3.

B. Testing of the Model

The model that has been formed is tested for its accuracy level by entering test data derived from training data. Because the data obtained in this study after the preprocessing process is only 136 data, therefore the cross validation method, Confusion Matrix, and the ROC Curve are used to test the level of accuracy. For the accuracy of the model for the artificial neural network algorithm is 89.89%.

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

1. Confusion Matrix

Table 1 is a confusion matrix for artificial neural network methods. It is known from 136 data, 69 classified as no (negative) predicted according to facts, then 7 data predicted no (negative) but facts are yes (positive), 53 data classified as yes (positive) predicted according to facts, and 7 data predicted yes (positive) but facts are no (negative).

Table 1. Confusion Matrix Model for Artificial Neural Network Method

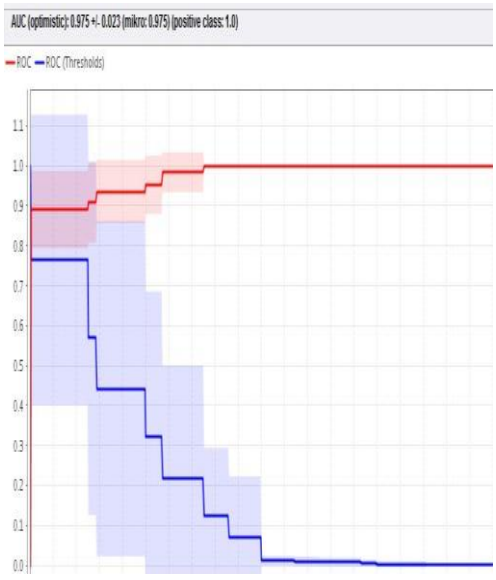
accuracy: 89.89% +/- 8.63% (mikro: 89.71%)

	true 0.0	true 1.0	class precision
pred. 0.0	69	7	90.79%
pred. 1.0	7	53	88.33%
class recall	90.79%	88.33%	

Source: Processing Results Using RapidMiner 5.3 (2019)

2. ROC Curve

The calculation results are visualized by the ROC curve. The ROC curve in Figure 4 expresses the confusion matrix. Horizontal lines are false positives and vertical lines true positives. The ROC curve for artificial neural network algorithms is shown in Figure 4 below



Source: Processing Results Using RapidMiner 5.3 (2019)

Figure 5 ROC Curve with Artificial Neural Network Method

From the picture above it can be seen that the undercurve area value (AUC) of the neural network method is 0.975.

For data mining classifications, the AUC value can be divided into several groups [5]:

0.90 - 1.00 = Exellent Clasification
0.80 - 0.90 = Good Clasification
0.70 - 0.80 = Fair Clasification
0.60 - 0.70 = Poor Clasification
0.50– 0.60 = Failure

Based on the grouping above, it can be concluded that the artificial neural network method is classified as very good (Exellent Clasification) because it has an AUC value between 0.90-1.00.

V. CONCLUSION

The conclusion that can be taken based on this research is that the performance of the artificial neural network algorithm model provides a level of truth accuracy of 89.89% with an area under the curve (AUC) of 0.975. This shows that the model including the classification category is very good (exellent clasification) because it has an AUC value between 0.90-1.00.

REFERENCE

[1] Alpaydin, E. (2010). Introduction to Machine Learning. London: The MIT Press.
[2] Amrin, A. (2018). Perbandingan Metode Neural Network Model Radial Basis Function Dan Multilayer Perceptron Untuk Analisa Risiko Kredit Mobil. *Jurnal Paradigma*, *XX*(1), 31–38. Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/2783
[3] Amrin, A., & Saiyar, H. (2018). Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes. *Jurna Jurikom*, *5*(5), 498–502. Retrieved from https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom/article/view/900/864
[4] Fine, J. (2012). *An Overview Of Statistical Methods in Diagnostic Medicine*. Chapel Hill.
[5] Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer.
[6] Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Soft Computing* (Vol. 54). San Fransisco: Morgan Kauffman. https://doi.org/10.1007/978-3-642-19721-5
[7] Kusumadewi, S. (2010). *Pengantar Jaringan Syaraf Tiruan*. Yogyakarta: Teknik Informatika FT UII.
[8] Larose, D. . (2005). Discovering Knowledge in Data. New Jersey: John Willey & Sons, Inc.
[9] Liao, T. W. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Application*. Singapore: World Scientific Publishing.
[10] Maimon, O., & Rokach, L. (2010). *Data Mining And Knowledge Discovery Handbook*. New York: Springer.

Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i2.10028
e-ISSN : 2541-2019
p-ISSN : 2541-044X

[11] Myatt, G. J. (2007). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey: John Wiley & Sons, Inc.
[12] Orhan, E., Temurtas, F., & Tanrıkulu, A. Ç. (2010). *Tuberculosis Disease Diagnosis Using Artificial Neural Networks*. Springer, 299-302.
[13] Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
[14] Sogala, S. S. (2006). *Comparing the Efficacy of the Decision Trees with Logistic Regression for Credit Risk Analysis*. India.
[15] Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. Berlin Heidelberg New York: Springer.
[16] Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.
[17] Widoyono. (2011). *Penyakit Tropis Epidemiologi, Penularan, Pencegahan dan Pemberantasan.* Jakarta: Erlangga.
[18] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning and Tools*. Burlington: Morgan Kaufmann.