Intelligent Diagnosis System for Acute Respiratory Infection in Infants

Subiyanto and Anggraini Mulwinda
Department of Electrical Engineering
Universitas Negeri Semarang
Semarang, Indonesia
subiyanto@mail.unnes.ac.id

Dwi Andriani
Computer and Network Engineering
SMK Gajah Mada Purwodadi
Purwodadi, Indonesia
dwi.anndriani@gmail.com

***Abstract*—Acute Respiratory Infections (ARI) became the main cause of morbidity and mortality of infectious diseases in the world. Recent studies have focused on the use of data mining techniques to build predictive models that are able to diagnose the ARI. The objective of this research is to develop a diagnosis system to predict ARI in infants using C4.5 algorithm. The algorithm used to build a decision tree. This research is a collaboration authors with the hospitals and doctors. The dataset was obtained from medical records of patients with respiratory disease from a hospital. The data are used as training data and test data. Symptoms that are used as input systems are the danger sign, fever, cough, shortness of breath and fast breathing. The first step is to pre-process subsequent data algorithm classification to form a decision tree. After the decision tree was formed, continued set the rules. That decision rules are implemented to establish the diagnosis system. Validation is done by comparing the results of diagnosis system with the doctor diagnosis. The comparison showed that the results of diagnosis system approaching the diagnosis of doctor. From these results, it can be concluded that the C4.5 algorithm could help to diagnose ARI. However, further investigation with the larger dataset is still needed.**

Keywords*—*Acute Respiratory Infections; prediction; diagnosis system; C4.5

I. Introduction

Acute Respiratory Infections (ARI) became the main cause of morbidity and mortality of infectious diseases in the world. Almost 4 million people die from ARDs each year, 98% caused by lower respiratory infections [1]. The mortality rate is very high in infants, children and the elderly, especially in countries with low and middle income per capita. Pneumonia, which is one type of ARI is the main killer of children under five in the world, more than the deaths due to AIDS, malaria, and measles. Early prediction of acute respiratory infections is one of the control measures to reduce the risk of transmission [2]. Thus, an urgent is need to diagnose ARI as early as possible.

Data mining is one technique that can be used to help establish a medical decision [3][4]. These techniques can effectively diagnose at an earlier stage by extracting valuable information from the patient dataset [3][4][5]. The most common method used in data mining techniques are neural networks, decision tree, apriori, regressions, k-means, Bayesian networks, etc [3]. Compared with the others, the method of decision tree is the fastest and most accurate [6][7]. The decision tree is a graphical representation which is described by the hierarchical tree. Decision-making with decision tree method is capable to produce high classification accuracy with a simple representation of the collected knowledge and very appropriate to support the decision-making process in medicine [4][8]. Some previous researchers have succeeded in applying decision tree method to solve various cases. Ramezankhani et al. (2015) have identified the condition of patients using decision tree method with fairly accurate results, which reached 90.5% [9].

Decision tree used a variety of machine-learning algorithms that can help to get information from large data [3]. Some of them are ID3, C4.5, and Naive Bayes algorithm. In recent years, the C4.5 algorithm is often used to help establish a medical decision. The C4.5 algorithm which is triggered by Quinlan is an extension of the ID3 algorithm [6][10][11]. The algorithm is widely used for the classification because of the fast, effective and produce high precision [12][13]. Some previous researchers have succeeded in applying the C4.5 algorithm to solve various cases. Hssina et al. were comparing the performance between ID3, C4.5, C5.0, and CART algorithms [11]. The results suggest that the C4.5 algorithm is the best to make a machine learning algorithm than others [11]. Aljaaf J. et al. managed to develop a predictive model capable of predicting the incidence of heart failure. This article used a multi-level risk assessment, in which five levels of risk of heart failure is predicted using the C4.5 algorithm. The proposed prediction models have a fairly high degree of accuracy, which reached 86.53% [14].

Moreover, Rajesh and Sheila implement the C4.5 algorithm to diagnose breast cancer. As a result, the algorithm shows the best results in the dataset and has the smallest error value [15]. Radha and Srinivasan were comparing the C4.5 algorithm, SVM, k-NN, PNN and BLR in predicting diabetes. The comparison showed that the C4.5 algorithm expressed as a learning algorithm that has the highest level of accuracy than other learning algorithms [16].

However, although the C4.5 algorithm already very commonly used in the health field such as to diagnose cancer and heart case disease, its application to diagnose of acute respiratory infections still rare. In this work, we investigate how decision tree based on C4.5 algorithm can help for the ARI disease prediction. The aim is to apply C4.5 algorithm to

978-1-5090-5864-8/17/$31.00 ©2017 IEEE

build prediction system for ARI. This system was developed using data mining, which includes the step of pre-processing, classification using the C4.5 algorithm, build the decision tree and generating the decision rules.

II. Intelligent Diagnosis System

A. *Classification by C4.5 Algorithm*

The C4.5 algorithm which is triggered by Quinlan is an extension of the ID3 algorithm [5][6][11]. Below is the algorithm to generate decision tree [8].

Input:

1) Training dataset D, which is a set of training observations and their associated class value.
2) Attribute list A, the set of candidate attributes.
3) Selected splitting criteria method.

Output: A decision tree.

In this paper, the following splitting criteria were investigated are entropy, information gain, and gain ratio. The gain ratio is used as the basis for selecting the attribute that is used as the root of the decision tree [17].

The Entropy defines as:

$$Entropy(D) = \sum_{i=1}^{n} -p_i * \log_2 p_i \tag{1}$$

where D is partition the data, n is the number of the target attribute value, p_i is probability value to the target attribute-i.

$$Entropy(A) = \sum_{i=1}^{n} -pA_i * \log_2 pA_i \tag{2}$$

where A is attribute value, n is the number of the target attribute value, pA_i is probability value of the target attribute i of the value of attribute A.

The Information gain is expressed as:

$$gain(D, A) = Entropy(D) - \sum_{i=1}^{n} \frac{|A_i|}{|D|} * Entropy(A_i) \tag{3}$$

where D is partition data, A is attributes, n is the number of input attribute value of attribute A, $\frac{|A_i|}{|D|}$ is probability attribute value input A to-i.

The Split Info is given by:

$$Split\ Info(Ti) = \sum_{i=1}^{n} -\left(\frac{Ti}{T}\right)^{*} \log_2\left(\frac{Ti}{T}\right) \tag{4}$$

where Ti is a number of cases the value of the variable, T is the number of total cases.

The gain ratio is defined as:

$$gain\ Ratio(A) = \frac{gain\ (A)}{SplitInfo(A)} \tag{5}$$

The attributes which have the highest ratio gain value used as a root node.

B. *Experimental Dataset*

This research was conducted with the hospitals and general doctors. Data collection was performed from medical records of patients with respiratory disease at the Dr. Adhyatma Tugurejo Semarang Hospital. The determination of the attributes based on the journal and guide books of ARI diagnoses in children who already consulted with the doctor concerned. Validation of the system will be performed by comparing the results of system diagnosis with the diagnosis the doctor.

The authors used the data 106 patients were each divided into groups of training data and test data, where 79 data for training data and 27 data for test data. Data from each patient distinguished by attributes the symptoms include danger sign, fever, cough, shortness of breath, and fast breathing [18]. The attributes of danger sign, fever, cough, shortness of breath and fast breathing have value yes and no. While diagnose have value severe disease, not pneumonia, pneumonia and severe pneumonia. The C4.5 algorithm was used in this study for the rules to identify ARI disease level. Respiratory infections are divided into four, namely (1) Severe Disease; (2) Not Pneumonia; (3) Pneumonia; and (4) Severe Pneumonia.

C. *Diagnosis Procedure of ARI*

Fig. 1 shows the flowchart of steps to build a diagnosis system of ARI using C4.5 algorithm. According to the flowchart above, stages in making a decision tree to build diagnosis system of ARI starts from the determination of training data and test data. After the training data set, the next step is to calculate the entropy of each attribute. From the calculation of the entropy, can be searched gain value, split info and gain ratio. An attribute with the highest gain ratio value serves as the root node. The process of calculating the entropy, gain, split info and gain ratio continues until an empty attribute. The results of the decision tree that are what will be used as the rules for the development of ARI diagnosis system.

The number of the target attributes value is denoted as i. So, for target attribute Severe Diseases is denoted as i=1, target attribute Not Pneumonia is denoted as i=2, target attribute Pneumonia is denoted as i=3 and target attribute Severe Pneumonia is denoted as i=4. So, based on the equation 1 wherein the amount of the value of a = 4, the equation for calculating the entropy in cases of ARI disease are as follows

$$Entropy(D) = \sum_{i=1}^{4} -p_i * \log_2 p_i \tag{6}$$

The attributes used in diagnosis system to help diagnose the ARI disease denoted as Ai. The attributes used are as follows:

A1 = the danger sign

A2 = fever symptom

A3 = cough symptom

A4 = shortness of breath

A5 = fast breathing

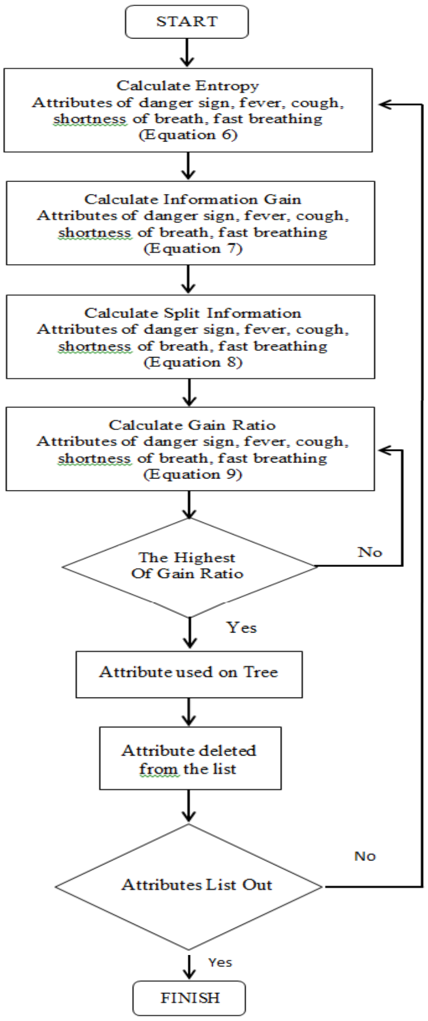
START
Calculate Entropy
Attributes of danger sign, fever, cough, shortness of breath, fast breathing
(Equation 6)
Calculate Information Gain
Attributes of danger sign, fever, cough, shortness of breath, fast breathing
(Equation 7)
Calculate Split Information
Attributes of danger sign, fever, cough, shortness of breath, fast breathing
(Equation 8)
Calculate Gain Ratio
Attributes of danger sign, fever, cough, shortness of breath, fast breathing
(Equation 9)
The Highest Of Gain Ratio
No
Yes
Attribute used on Tree
Attribute deleted from the list
Attributes List Out
No
Yes
FINISH

Fig. 1. Flowchart of C4.5 algorithm on diagnosis system of ARI

TABLE I. DATASET DISTRIBUTION OF TRAINING DATA

SYMPTOMS		NUMBER OF CASE	A	B	C	D
TOTAL		79	11	39	7	22
Danger sign	Yes	11	11	0	0	0
	No	68	0	39	7	22
Fever	Yes	57	10	30	3	14
	No	22	1	9	4	8
Cough	Yes	57	7	24	7	19
	No	22	4	15	0	3
Shortness of breath	Yes	22	1	4	2	15
	No	57	10	35	5	7
Fast breathing	Yes	11	2	3	5	1
	No	68	9	36	2	21

[a.] Note: A = Severe Disease; B = Not Pneumonia; C = Pneumonia; D = Severe Pneumonia

Based on the equation 3 wherein the amount of the value of a= 4, that is severe disease, not pneumonia, pneumonia and severe pneumonia, the equation for calculating the entropy in cases of ARI disease are as follows

$$gain(D,A) = Entropy(D) - \sum_{i=1}^{4} \frac{|A_i|}{|D|} * Entropy(A_i) \quad (7)$$

After determining the entropy and gain, the next step is to calculate the value of split info. According to equation 4, the split info can be calculated using the following equation.

$$SplitInfo(Ti) = \sum_{i=1}^{4} - \left(\frac{Ti}{T}\right)^{*} \log_2 \left(\frac{Ti}{T}\right) \quad (8)$$

where Ti is a number of cases the value of the variable, T is the number of total cases. The gain ratio is defined as:

$$gain\ Ratio(A) = \frac{gain\ (A)}{SplitInfo(A)} \quad (9)$$

The process of calculating the entropy, gain, split info and gain ratio continues until an empty attribute.

III. Result and Discussion

To build decision trees, the steps should be done is to calculate the value of entropy, information gain, split info and gain ratio. The calculation is performed until the data cannot be partitioned again. After all the calculations are done and the training data cannot be partitioned again, a decision tree can be shaped to provide a prediction of patient's illness. From these results, the final decision tree can be described as Fig. 2.

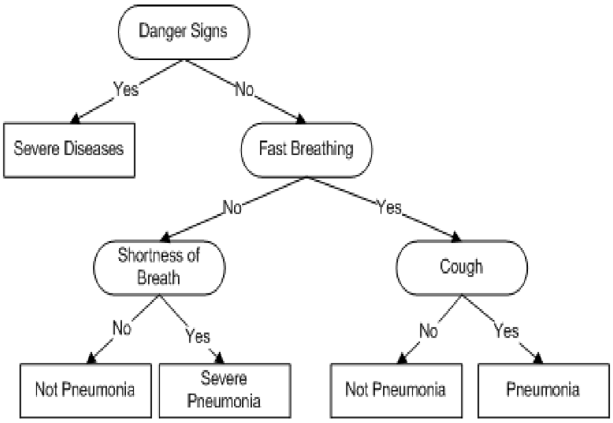

Fig. 2. Final Decision Tree

Rules resulting from the decision tree is then implemented on ARI diagnosis system, is as follows:

1. IF Danger Signs = Yes THEN Severe Diseases.
2. IF Danger Signs = No AND Fast Breathing = No AND Shortness of Breath = No THEN Not Pneumonia.
3. IF Danger Signs = No AND Fast Breathing = No AND Shortness of Breath = Yes THEN Severe Pneumonia.
4. IF Danger Signs = No AND Fast Breathing = Yes AND Cough = No THEN Not Pneumonia.
5. IF Danger Signs = No AND Fast Breathing = Yes AND Cough = Yes THEN Pneumonia.

System testing is performed to determine the accuracy of the output system. This level of accuracy is obtained by comparing the results of the doctor diagnosis with the diagnosis system obtained by inserting the symptoms experienced by patients with the system. Details of the test result on the system shown in Table II.

TABLE II. COMPARISON OF THE DOCTOR DIAGNOSIS WITH THE OUTPUT OF DIAGNOSIS SYSTEM

No.	Doctor Diagnosis	Output of Diagnosis System
1.	Severe Diseases	Severe Diseases
2.	Severe Diseases	Severe Diseases
3.	Severe Diseases	Severe Diseases
4.	Severe Diseases	Severe Diseases
5.	Not Pneumonia	Not Pneumonia
6.	Not Pneumonia	Not Pneumonia
7.	Not Pneumonia	Not Pneumonia
8.	Not Pneumonia	Not Pneumonia
9.	Not Pneumonia	Not Pneumonia
10.	Not Pneumonia	Not Pneumonia
11.	Not Pneumonia	Not Pneumonia
12.	Not Pneumonia	Not Pneumonia
13.	Not Pneumonia	Not Pneumonia
14.	Not Pneumonia	Severe Pneumonia
15.	Not Pneumonia	Severe Pneumonia
16.	Not Pneumonia	Not Pneumonia
17.	Not Pneumonia	Not Pneumonia
18.	Pneumonia	Not Pneumonia
19.	Pneumonia	Pneumonia
20.	Severe Pneumonia	Severe Pneumonia
21.	Severe Pneumonia	Severe Pneumonia
22.	Severe Pneumonia	Severe Pneumonia
23.	Severe Pneumonia	Severe Pneumonia
24.	Severe Pneumonia	Severe Pneumonia
25.	Severe Pneumonia	Pneumonia
26.	Severe Pneumonia	Not Pneumonia
27.	Severe Pneumonia	Severe Pneumonia

Any tests were conducted by the method "using test set" that with this method, decision tree obtained will be tested using test data. To find out how well the ability classifier, used confusion matrix. The confusion matrix is a method used to perform calculations on the accuracy of data mining concepts. So, the confusion matrix of the diagnosis system can be obtained in Table III.

TABLE III. CONFUSION MATRIX OF DIAGNOSIS SYSTEM

Diagnosis		System				Total
		A	*B*	*C*	*D*	
Doctor	***Severe Disease***	4	0	0	0	4
	Not Pneumonia	0	11	0	2	13
	Pneumonia	0	1	1	0	2
	Severe Pneumonia	0	1	1	6	8
Total		4	13	2	8	27

[b] Note: A = Severe Disease; B = Not Pneumonia; C = Pneumonia; D = Severe Pneumonia

From the results in Table III showed the correct diagnose of system and in accordance with the doctor's diagnosis as much as 22 data. Results correct diagnose of severe diseases as much as 4 data, correct diagnose of not pneumonia as much as 11 data, correct diagnose of pneumonia as much as 1 data, and correct diagnose of severe pneumonia as much as 6 data. This means that of the 27 data used as the test data, 22 data are expressed in accordance with the doctor diagnosis and the remaining 5 data that does not fit. If these results are calculated using the confusion matrix, it will get the value of 81,48% accuracy.

IV. CONCLUSION

In this paper, ARI diagnosis system built by applying the C4.5 decision tree algorithm. Preprocessing steps include:

integration phase, data cleaning and data transformation to produce a clean dataset that can be used in the next stage of data mining. The next step is classification by C4.5 algorithm, the following splitting criteria were investigated are entropy, information gain, split info, and gain ratio. The gain ratio is used as the basis for selecting the attribute that is used as the root of the decision tree. After all the calculations completed and a decision tree has been formed, the next step is to determine the rules of the decision. The decision rule is used as the basis for the decision system. Techniques validation is performed by comparing the output of the system with the result of the diagnosis by a doctor. In testing it was found that of 27 patients who used the data as test data, 22 data is identified in accordance with the doctor diagnosis and 5 data are not appropriate. Based on the research that has been done, it can be concluded that C4.5 algorithm showed 81,48% according to the doctor diagnosis, so it can be used to build the diagnosis system of ARI disease. However, further investigation with larger datasets is still needed to improve the accuracy of the system.

Acknowledgment

This research was supported by Department of Electrical Engineering, Universitas Negeri Semarang and the UEESRG. Authors are grateful for their cooperation and support.

References

[1] WHO 2004, ‘World Health Report 2004: Changing History’, WHO Library Cataloguing-in-Publication Data, pp. 120-121

[2] WHO 2007, ‘Infection prevention and control of epidemic-and pandemic-prone acute respiratory diseases in health care’, WHO Interim Guidelines, p.12.

[3] Mahmud, SA, & Ismail, M 2015, ‘Application of data mining in medical decision support system’, International Journal of Information System and Engineering, vol. 1, no. 1, p. 2.

[4] Podgorelec, V, Kokol, P, & Stiglic B 2002, ‘Decision Trees : an overview their use in medicine’, Journal of Medical Systems, Kluwer Academic/Plenum Press, vol. 26, no. 5.

[5] Minegishi, T, Ise, M, Niimi, A & Konishi, O 2009, ‘Extension of decision tree algorithm for stream data mining using real data’, Fifth International Workshop on Computational Intelligence & Applications, Chapter 208-2011.

[6] Teli, S & Kanikar P 2015, ‘A survey on decision tree based approaches in data mining’, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 4.

[7] Adeyemo, O, Adeyeye, T & Ogunbiyi D 2015, ‘Comparative study of ID3/C4.5 decision tree and multilayer perceptron algorithms for the prediction of typhoid fever’, African Journal of Computing & ICT, vol. 8, no. 1.

[8] Karaolis, M, Moutiris, J, Hadjipanayi, D & Pattichis, C 2010, ‘Assessment of the risk factors of coronary heart events based on data mining with decision trees’, IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 3.

[9] Ramezankhani, A, Pournik, O, Shahrabi, J, Khalili, D, Azizi, F & Hadaegh F 2015, Applying decision tree for identification of a low risk population for type 2 diabetes, tehran lipid and glucose study’, Research Article 2016, vol. 2636390, no. 1.

[10] Masetic, Z & Subasi, A, 2013, ‘Detection of congestive heart failures using C4.5 decision tree’, Europe Journal of Soft Computing, vol. 2, no. 2.

[11] Hssina, B, Merbouha, A, Ezzikouri, H & Erritali, M, ‘A comparative study of decision tree ID3 and C4.5’, International Journal of Advanced Computer Science and Applications.

[12] Sharma, S, Agrawal, J & Sharma, S 2013, ‘Classification through machine learning technique : C4.5 algorithm based on various entropies’, International Journal of Computer Applications, vol. 82, no. 16, p.1.

[13] Chauhan, H & Chauhan, A 2013, ‘Implementation of decision algorithm C4.5’, International Journal of Scientific and Research Publications, vol. 3.

[14] Aljaaf, J, Al-Jumeily, D, Hussain, A, Dawson, T, Fergus, P & Al-Jumaily, M 2015, ‘Predicting the likelihood of heart failure with a multi level risk assessment using decision tree’ Proceedings of The Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering

[15] Rajesh, K & Anand, S 2012, ‘Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm’, International Journal of Advanced Research in Computer and Communication Engineering, vol. 1, no. 2.

[16] Radha, P & Srinivasan, B 2014, ‘Predicting diabetes by cosequencing the various data mining classification techniques’, International Journal of Innovative Science, Engineering & Technology (IJISET), vol. 1, no. 6.

[17] Han, J, Kamber & Pei 2012, Data Mining Concepts and Techniques. 3rd ed., Elsevier Inc. San Francisco.

[18] Demmler, G, MD & Ligon, B 2003, ‘Severe acute respiratory syndrome (SARS): a review of the history, epidemiology, prevention and concerns for the future’, Seminars in Pediatric Infectious Diseases, vol. 14, no. 3, p. 244.