ReliefF Feature Selection and Bayesian Network Model for Hepatitis Diagnosis

Fetty Tri Anggraeny[1], Intan Yuniar Purbasari[2], Evi Suryaningsih[3]
Informatics, Department of Computer Science, UPN “Veteran” Jawa Timur[1,2,3]
fettyanggraeny.if@upnjatim.ac.id[1], intanyuniar.if@upnjatim.ac.id[2],
evi.surya20@gmail.com[3]

ABSTRACT

A doctor diagnose a disease by evaluating patient condition or by comparing with another patient that have similar conditions or symptoms. In computer science, this task can be done by a computer program that included intelligent algorithm in it. Some disease have similar symptoms, such as typhoid fever, hepatitis, and dengue fever. Based on UCI database there are 17 symptoms of Hepatitis that may be similar with other disease, so it needs a method to find the major symptoms. In this research, we proposed hepatitis diagnose using statistic Bayesian network and find major symptoms using ReliefF algorithm. ReliefF algorithm resulting 4 majority symptoms and used to constructing Bayesian Network. ReliefF and Bayesian Network have 76,8% accuracy, 76,5% precision, and 100% recall for 69 test data.

Keywords: Hepatitis, ReliefF, Bayesian Network, Probabilistic

1. Introduction

Hepatitis is a common term that means liver inflammation. "Hepa" means a link with the liver, while "itis" means inflammation (such as in atritis, dermatitis, and pancreatitis). Hepatitis has several causes, including: toxins and chemicals such as excess alcohol; diseases that cause the immune system to attack healthy tissues in the body, called autoimmune diseases; and microorganisms, including viruses [1].

Based on data from the Ministry of Health of the Republic of Indonesia in the 2014 Data and Information Center document, Indonesia is a country with high endemicity of hepatitis B, the second largest in Southeast Asia after Myanmar. It is estimated that there are currently 28 million Indonesians infected with Hepatitis B and C, 14 million of whom are potentially chronic, and of which 1.4 million are chronically potentially liver cancer [2].

There have been many attempts by the government to reduce the number of hepatitis patients, such as immunization, global hepatitis day warning, seminars related to hepatitis, and early detection. Early detection is done to pregnant women and health workers. Early detection in pregnant women to bypass the chain of vertical hepatitis transmission, while early detection of health workers to prevent horizontal transmission.

Based on the above conditions, in this paper will be modeled a learning machine capable of detecting people with chronic hepatitis disease. Similar research has been done by several researchers using different methods. Learning machines that have been used include neural network [3-7], decision tree [8], SVM [9], and Naive Bayes [4]. All of the predecessor's studies provide good results with an accuracy level above 80%.

Some studies that use Bayesian Network in different cases provide good results. Fallahi et al [10] with the title Expert System To Detect Breast Cancer Using Preprocessing Data and Bayesian Network. Based on these research is done preprocessing on the data to solve the problem of unbalanced data and lost data. Then used ReliefF Algorithm to reduce the number of database features and classification with bayesian network. The results show that the use of ReliefF Algorithm method with bayesian network can yield better result for data classification and get significant accuracy to diagnose breast cancer disease. Research conducted by Rizki et al [11] with the title of Multinomial Bayesian Network Model on Rainfall Simulation Data. The Bayesian Network structure is built using the K2 algorithm.

In this study we propose to model the diagnosis of hepatitis disease using Bayesian

Network and to know the symptoms of hepatitis which have high factor used ReliefF algorithm.

2. Research Method

Figure 1 shown the methods of this research. The diagnostic process begins by selecting dataset features using WEKA 3.6.12 (Waikato Environment for Knowledge Analysis) application tools. The dataset feature is selected using the ReliefF Algorithm and generates the ranking and weight of each feature. The results of the selected features will be selected based on the largest weights. The result of feature selection using ReliefF can be seen in Table 2 and 4 features with the highest ranking can be seen in Table 3. The selected features are then used to build bayesian network.

A. Hepatitis Dataset

Hepatitis disease dataset obtained from UCI Machine Learning Repository [12]. The data used as many as 155 records that have 19 attributes and 1 class attribute. Of the 155 record data, 32 records labeled die and 123 records labeled live. The original hepatitis disease dataset contains many missing values, data with empty attributes of 75 records and attributes with complete data totaling 80 records. Of 80 data records, 13 records labeled die and 67 records labeled live. Description of the attribute of the hepatitis dataset is described in Table 1.

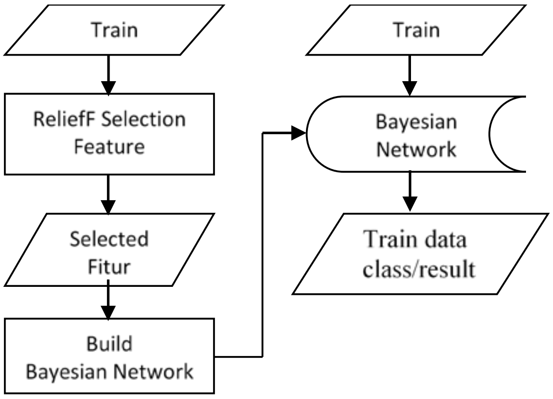


Figure 1. Methodology Research.

Tabel 1. Information Attribute Of Hepatitis Dataset

#	Atribut	Value
1	*Age*	Continue
2	*Sex*	1= Laki–Laki, 2= Perempuan
3	*Steroid*	1 = *No*, 2 = *Yes*
4	*Antivirals*	1 = *No*, 2 = *Yes*
5	*Fatigue*	1 = *No*, 2 = *Yes*
6	*Malaise*	1 = *No*, 2 = *Yes*
7	*Anorexia*	1 = *No*, 2 = *Yes*
8	*Liver Big*	1 = *No*, 2 = *Yes*
9	*Liver Firm*	1 = *No*, 2 = *Yes*
10	*Spleen Palpable*	1 = *No*, 2 = *Yes*
11	*Spiders*	1 = *No*, 2 = *Yes*
12	*Ascites*	1 = *No*, 2 = *Yes*
13	*Varices*	1 = *No*, 2 = *Yes*
14	*Bilirubin*	0.3 until 7.6
15	*Alk Phosphate*	30 until 295
16	*Sgot*	14 until 648
17	*Albumin*	2.4 until 6.4
18	*Protime*	0 until 100
19	*Histology*	1 = *No*, 2 = *Yes*

B. ReliefF Feature Selection

ReliefF [13] is a classical feature selection algorithm. It utilizes the correlation between the characteristics to make similar samples close and keep heterogeneous samples apart in order to achieve the purpose of the feature selection. ReliefF algorithm is the development of Relief algorithm that is not able to overcome inclomplete data and only limited to 2 class problem only. ReliefF algorithm is made to solve the problems that can not be overcome by Relief algorithm.

ReliefF algorithm is used to solve single label

problem [14]. Assume that there are *n* instances and *L* labels. Let $P \in R^f$ be the full set of features, $p \in P$ be a feature, $X=[x_1, x_2, ..., x_n] \in R^{n \times f}$ denote instances and let $Y =[y_1, y_2, ..., y_n] \in R^{n \times L}$ denote the instances with labels. One instance represented by $x_i \in R^f$ can be expressed as x_i=[$p_i^1, p_i^2, \ldots, p_i^f$]. It is associated with a set of labels by a binary vector y_i={0,1}^{l} , and y_i(l)=1 if x_i belongs to the *l* th class and y_i(l)=0 otherwise. Since an instance owns multiple labels, $\sum y_i(\text{l}) \geq 1$.

For the classical ReliefF [13, 15], the algorithm samples *m* instances randomly from the dataset. For each sample point x_t (1≤*t*≤*n*), it finds *K* nearest neighbors that belongs to the same class *C* as x_t named as Hit and for other (*L*–1) classes (other than *C*), it also finds *K* nearest neighbors denoted as Miss (*C*); So the formula for updating every feature is computed as,

$$W_P = W_P - \sum_{j=1}^{K} \frac{d(p, x_t, H_j)}{m \cdot K} +$$

$$\sum_{C \neq C(x_t)} \sum_{j=1}^{K} \frac{P(C)}{1-P(C(x_t))} \cdot \frac{d(p, x_t, M_j)}{m \cdot K} \qquad (1)$$

Where W_P denotes the value of feature *p*, *P*(*C*) is the priori probability of the label class *C*, and $d(p, x_t, x_j)$ is the distance between x_t and x_j on feature *p* (usually the Euclidian distance).

WEKA 3.6.12 application tools is used to implement ReliefF Algorithm selection features. Eighty training data on hepatitis disease were entered into WEKA application with two class labels: live and die. Where 67 records are labeled live and 13 records are labeled die. The ranking of feature selection on 19 attributes of hepatitis dataset using WEKA application can be seen in Table 2. Table 3, a feature with a high role in the classification of having a weight value greater than 0.2 is used to build a bayesian network.

Table 2. Selection ranks features of hepatitis dataset.

Ranked	Weight	Attributes
1	0.3825	histology
2	0.30375	ascites
3	0.2525	spiders
4	0.25	malaise
5	0.18125	varices
6	0.16375	fatigue
7	0.155	liver_firm
8	0.145	steroid
9	0.1225	spleen_palpable
10	0.10621	albumin
11	0.10069	bilirubin
12	0.0975	antivirals
13	0.08991	protime
14	0.06771	alk_phosphate
15	0.065	anorexia
16	0.04625	liver_big
17	0.03043	age
18	0.02125	sex
19	-0.00505	sgot

Table 3. Selected feature.

Ranked	Weight	Attributes
1	0.3825	histology
2	0.30375	ascites
3	0.2525	spiders
4	0.25	malaise

C. Bayesian Network

Bayesian network is a graphical model to represent interactions between variables based on bayes theorem. Bayes's theorem is an approach to uncertainty as measured by probability. The Bayes Theorem formula is given in equation 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

P(A|B) = posterior probability, ie A chance occurs after B happen.

P(B|A) = likelihood, ie B chance occurs after A occurs.

P(A) = prior, the probability of occurrence A

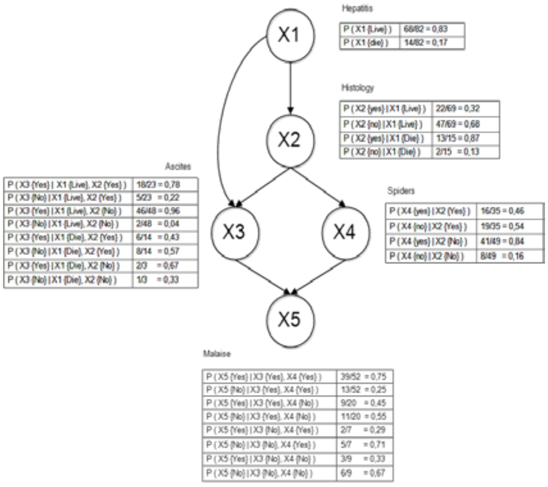
P(B) = Opportunity occurrence B

Bayesian network is one of the simplest Probabilistic Graphical Model (PGM) constructed from probabilistic theory and graph theory. The probabilistic theory is directly related to the data whereas the graph theory is directly related to the form of representation to be obtained. Bayesian networks can represent causal relationships among variables contained in the bayesian network structure. For example, a bayesian network may represent a probabilistic relationship between disease and symptoms. Bayesian networks can be used to calculate the probability of the presence of various symptoms of the disease.

Bayesian networks can make probabilistic inference. The probabilistic inference is to predict the value of unknown variables directly using the values of other known variables [16]. Probabilistic inference can be done if Joint Probabillity Distribution (JPD) of all known variables is known [16]. JPD is the probability of all occurrences of variables occurring simultaneously. Probabilistic inference can be done if the bayesian network has been built, so what needs to be done first is to build the Bayesian Network structure [17].

```
procedure K2;
{Input: A set of n nodes, an ordering on
the nodes, an upper bound u on the
number of parents a node may have, and
a database D containing m cases.}
{Output: For each node, a printout of the
parents of the node.}
for i := 1 to n do
   $\pi i := \emptyset$;
   Pold := f (i,$\pi i$);{This function is
computed using Equation 2}
```

Figure 2. K2 algorithm [18].



Hepatitis

P (X1 {Live})	68/82 = 0,83
P (X1 {die})	14/82 = 0,17

Histology

P (X2 {yes} | X1 {Live})	22/69 = 0,32
P (X2 {no} | X1 {Live})	47/69 = 0,68
P (X2 {yes} | X1 {Die})	13/15 = 0,87
P (X2 {no} | X1 {Die})	2/15 = 0,13

Ascites

P (X3 {Yes} | X1 {Live}, X2 {Yes})	18/23 = 0,78
P (X3 {No} | X1 {Live}, X2 {Yes})	5/23 = 0,22
P (X3 {Yes} | X1 {Live}, X2 {No})	46/48 = 0,96
P (X3 {No} | X1 {Live}, X2 {No})	2/48 = 0,04
P (X3 {Yes} | X1 {Die}, X2 {Yes})	6/14 = 0,43
P (X3 {No} | X1 {Die}, X2 {Yes})	8/14 = 0,57
P (X3 {Yes} | X1 {Die}, X2 {No})	2/3 = 0,67
P (X3 {No} | X1 {Die}, X2 {No})	1/3 = 0,33

Spiders

P (X4 {yes} | X2 {Yes})	16/35 = 0,46
P (X4 {no} | X2 {Yes})	19/35 = 0,54
P (X4 {yes} | X2 {No})	41/49 = 0,84
P (X4 {no} | X2 {No})	8/49 = 0,16

Malaise

P (X5 {Yes} | X3 {Yes}, X4 {Yes})	39/52 = 0,75
P (X5 {No} | X3 {Yes}, X4 {Yes})	13/52 = 0,25
P (X5 {Yes} | X3 {Yes}, X4 {No})	9/20 = 0,45
P (X5 {No} | X3 {Yes}, X4 {No})	11/20 = 0,55
P (X5 {Yes} | X3 {No}, X4 {Yes})	2/7 = 0,29
P (X5 {No} | X3 {No}, X4 {Yes})	5/7 = 0,71
P (X5 {Yes} | X3 {No}, X4 {No})	3/9 = 0,33
P (X5 {No} | X3 {No}, X4 {No})	6/9 = 0,67

Figure 3. Bayesian network using 4 major attribute of hepatitis dataset.

K2 Algorithm

Bayesian network structure development is done manually with K2 algorithm. The K2 algorithm determines the Bayesian (Bs) network structure that maximizes the chances of Bs with Databases (Db). Observations are denoted by P (Bs, Db) by assuming that the variables are sorted. This algorithm begins by assuming that each node does not have a parent, then a parent is added where the addition increases the chance of the end result of the structure. If the addition of the parent no longer increases the chances of the end result of the structure, then the parent addition is stopped [11]. The pseudocode of K2 algorithm [18] developed by Cooper and Herzkowits.

The result of bayesian network development on four features that have been selected in table 4 are shown in figure 3.

3. Discussion.

Tests conducted on 69 data testing, 52 live category data and 17 data categories die. Based on the experiments conducted and made a comparison of real data, then of course there are some differences. The difference or error will be calculated error value. The value of this error will determine the quality of the created application. Testing result of 69 data is confussion matrix shown in Table 4.

Table 4. Confussion Matrix Of Testing Data.

Predict / Real	Die	Live
Die	1	16
Live	0	52

$$Accuracy = \frac{1+52}{1+16+0+52} \times 100\% = \frac{53}{69} = 76{,}8\ \%$$

$$Precision = \frac{52}{52+16} \times 100\% = \frac{52}{68} = 76{,}5\ \%$$

$$Recall = \frac{52}{52+0} \times 100\% = \frac{52}{52} = 100\%$$

Based on the result of accuracy, precision, and recall from matrix confusion in table 4, the accuracy of hepatitis diagnosis application program is 76,8%, precision or level of accuracy between information requested by user and answer given by system equal to 76,5% , and recall or system success rate in rediscovering an information of 100%. The difference in outcome of the trial is influenced by the number of attributes used because the results of the real data use 19 attributes while in the diagnostic program using 4 attributes.

4. Conclusion.

Based on experiment, we can get conclusion that implementation of REliefF as feature selection and Bayesian Network as decision method resulting experiments accuracy is 76.8%, the precision is 76.5%, and the recall is 100%.

In the future this research can be developed by making it into a ready-made applications, such as android applications. With the application then closer to the user in need. In terms of method development, it can use a classification method that is capable of processing data with missing value, such as decision tree.

Bibliographies

[1] C.W. Green, “Seri Buku Kecil: Hepatitis Virus and HIV,” Yayasan Spiritia, 2005.

[2] Infodatin, “Situasi dan Analisis Hepatitis,” The Ministry of Health of the Republic of Indonesia, 2014.

[3] G.S. Uttreshwar and A.A. Ghatol, “Hepatitis B Diagnosis Using Logical Inference and Self-Organizing Map,” Journal of Computer Science, Vol. 4, No. 12, 2008, pp. 1042-1050.

[4] B. Karlik, “Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers,” Journal of Science and Technology, Vol. 1, No. 1, 2011.

[5] O. Cetin, F. Temurtas, and S. Gulgonul, “An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function,” Dicle Medical Journal, Vol.42, No.2, 2015, pp. 150-157.

[6] B.S. Alshamrani and A.H. Osman, “Investigation of Hepatitis Disease Diagnosis using Different Types of Neural Network Algorithms,” IJCSNS International Journal of Computer Science and Network Security, Vol. 17, No. 2, February 2017.

[7] D. Panchal and S. Shah, “Artificial Intelligence Based Expert System For Hepatitis B Diagnosis,” International Journal of Modeling and Optimization, Vol. 1, No. 4, October 2011.

[8] V.S. Sowmien, V. Sugumaran, C.P. Karthikeyan and T.R. Vijayaram, “Diagnosis of Hepatitis using Decision tree algorithm,” International Journal of Engineering and Technology (IJET), Vol. 8, No. 3, June-July 2016, pp. 1414-1419.

[9] M.H. Afif, A.R. Hedar, T.H.A. Hamid and Y.B. Mahdy, “SS-SVM (3SVM): A New Classification Method for Hepatitis Disease Diagnosis,” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 2, 2013.

[10] Fallahi, Amir and S. Jafari, “An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network,” International Journal of Advanced Science and Technology, 34, 2011.

[11] Rizki, A. Nanda, Syaripuddin, and S. Wahyuningsih, “Model Multinomial Bayesian Network pada Data Simulasi Curah Hujan,” Statistika, Vol.12, No.2, 2012.

[12] G. Gong, “UCI Machine Learning Repository Hepatitis Dataset,” https://archive.ics.uci.edu/ml/datasets/Hepatitis. Accessed on August 18, 2015.

[13] I. Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” European conference on machine learning, April 1994, pp. 171-182.

[14] Y. Cai, M. Yang, and H. Yin, “Relieff-based multi-label feature selection,” International Journal of Database Theory and Application, Vol. 8, No. 4, 2015, pp.307-318.

[15] R. Durgabai, “Feature Selection using ReliefF Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 10, 2014.

[16] P.J. Krause, “Learning Probabilistic Networks,” Philips research laboratories, 1998.

[17] I. W. Santika, “Pengembangan Sistem Pakar Konsultasi Hama Dan Penyakit Tanaman Jeruk Menggunakan Metode Bayesian Network Berbasis Web,” Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika, Vol. 1, No. 4, 2012.

[18] C. Ruiz, “Illustration of the K2 Algorithm for Learning,” Lecture Notes on Machine Learning, Department of Computer Science, Worcester Polytechnic Institute, Massachusetts, 2005.