Comparative Study of Heart Disease Diagnosis Using Top Ten Data Mining Classification Algorithms

I Ketut Agung Enriko
PT Telkom Indonesia
Digital Service Division
Jakarta, Indonesia
+62 21 3860500
enrico@telkom.co.id

ABSTRACT

Data mining has been used for many purposes, especially for prediction system. In healthcare, data mining algorithms often used in disease diagnosis. Meanwhile, heart disease is known as a primary cause of death over the years. Many studies have been performed in heart disease diagnosis using data mining methods. There are some popular data mining algorithms that can be used in heart disease diagnosis, for example, k-Nearest Neighbor, CART, and AdaBoost. The algorithms are used to analyze a sample of cardiovascular patients data and predict the heart disease type that they suffer. Some parameters are taken from the patient, including EKG morphology, blood pressure, and information about the existence of chest pain, shortness of breath, palpitation, and cold sweat. In this study, medical records data are collected from Harapan Kita Hospital and utilized as a dataset sample in this research. Top ten data mining classification algorithms are used in diagnosing heart disease from Harapan Kita Hospital data and examining their performance by checking the accuracy and speed.

CCS Concepts

• **Information systems~Nearest-neighbor search**

Keywords

k-Nearest Neighbor; data mining; machine learning; machine learning algorithms

1. INTRODUCTION

These years, data mining has been implemented in many sectors including many areas in industries, engineering, and science. In telecommunication industry, for example, data mining can be used in analyzing churn pattern so the provider can avoid losing their customers with offering accurate promotions [1]. In the retail business, data mining role is in analyzing customer behaviour and products connection information so the arrangement of store products can be optimized [2]. The Internet industry uses data mining to analyze web browsing patterns to find out the best website design [3]. Even, data mining has been utilized in financial industry for fraud detection purpose [4].

Moreover, healthcare or medical is one of the sectors which data mining is increasingly applied, especially for clinical decision purpose. Data mining can help in predicting patient disease using patient's information related to his/her health. Various data mining algorithms can be used in medical prediction system, for example, Naive Bayes, Decision Tree, Random Forest, k-Nearest Neighbor (kNN), and Support Vector Machine (SVM). Previous studies have been performed in data mining implementation for medical diagnosis. A study [5] uses data mining techniques for identifying diabetic patients. Another study [6] uses data mining techniques for analyzing stroke severity index. Then, [7] investigates the data mining for cancer detection using serum proteomic profiling.

Meanwhile, heart disease or cardiovascular disease (CVD) has known as a dangerous disease. It is reported that CVD causes about thirty percents of death in the world [8,9,10]. There are many research related to heart disease prediction system using data mining techniques. A study [11] predicts heart diseases using Bagging Algorithm. The authors use a dataset from University of California, Irvine (UCI) KDD Archive to predict whether a patient will have a heart disease or not. They found that using Bagging Algorithm; the prediction accuracy is 81.41%, a better result comparing to Decision Tree which gave 78.91% accuracy.

The research from Kumari and Godara [12] compares some data mining algorithms in heart disease diagnosis. They use UCI repository, Cleveland cardiovascular dataset for the analysis. Their conclusion is the comparison table of algorithm prediction accuracy as follows: Ripper 81.08%; Decision Tree C4.5 79.05%; Artificial Neural Networks-Multilayer Perceptron (ANN-MLP) 90.06%; and SVM 84.12%.

Then a recent research [13] proposes heart disease diagnosis system using kNN algorithm, with the data of UCI, Hungarian Dataset. They aimed to simplify the number of parameters used, from 14 to 8, for some reasons. The accuracy of kNN algorithm is about 82%.

This paper performs a comparative study of heart disease diagnosis system using top ten data mining classification algorithms [14]. Six data mining classification algorithms are discussed (C4.5, SVM, Ada Boost, kNN, Naive Bayes, and CART). The four others are Random Forest, Bagging Algorithm, Logistic Regression, and Multilayer Perceptron (MLP). The novelty of this study is that real medical data from a hospital are used. The data collected from Harapan Kita Hospital (HKH), an educational cardiovascular hospital located in Jakarta. There are 450 data that collected for this research purpose.

ICFET 2019, June 1–3, 2019, Beijing, China

ACM ISBN 978-1-4503-6293-1/19/06…$15.00
https://doi.org/10.1145/3338188.3338220

This paper is organized as follows. A literature review about top ten data mining classification algorithms is presented in Section 2. The materials and methodology are written in Section 3. Result and discussion is the topic of Section 4, and the conclusion of this study is contained in Section 5.

2. LITERATURE REVIEW

There are well-known data mining algorithms usually used in many fields. Ten of them are discussed in this paper as follows.

2.1.C4.5

C4.5 is an algorithm used to build a decision tree [15]. This algorithm is an improvement from prior works: CLS [16] and ID3 [17]. C4.5 works by growing initial tree with the divide-and-conquer algorithm. The first step is, if all cases in given cases set S belong to the same class, or S is small, the tree is a leaf named with the most frequent class in S. If it is not, then a test should be chosen refers to a single attribute with two or more outcomes. This test will be the root of the tree with a branch for each outcome. Then S should be partitioned into corresponding subsets S1, S2, and so on according to each case's outcome. The same procedure has to be applied repeatedly to each subset.

While C4.5 has a good accuracy in prediction result, it has some downsides, for example, requiring a large amount of memory and CPU time [14].

2.2. SVM

SVM [18] is a robust and accurate algorithm with strong theoretical foundation, requires less training data, and insensitive [14]. SVM works by finding the best classification function to differ between members of two classes in the training data, with geometrical approach. A hyperplane f(x) is used to separate dataset into two classes. The mathematical model of SVM is written as follows:

$$L_p = \frac{1}{2}|\vec{w}| - \sum_{i=1}^{t} \alpha_i y_i (\vec{w} \cdot \vec{x_i} + b) + \sum_{i=1}^{t} \alpha_i \qquad (1)$$

Where t is the number of training data, αi (i=1,...,t) are non-negative figures that the derivatives of Lp with respect to αi are 0. Lp is the Lagrangian, and αi are the Lagrange multipliers. The vector ŵ and constant b determine the hyperplane.

Despite its popularity of robustness, SVM has certain drawbacks, for example, if it should handle the training data are not linearly separable [14].

2.3. AdaBoost

AdaBoost [19] is a type of ensemble method algorithm which has a strong theoretical foundation, simple, and promises a very good prediction. Ensemble learning is the method which uses multiple learners in problem-solving [20]. Usually, the performance of multiple learners is better than a single learner, so ensemble method is a great choice. The pseudo-code of AdaBoost is depicted in Figure 1.

AdaBoost has been successfully implemented in various fields, mainly focused on computer vision or face detection system [14].

2.4. kNN

kNN [21,22] is a classification method which works by finding a group of k objects in the training data that are nearest to a test datum. A set of training data, a distance of similarity, and a value of k (the number of nearest neighbors) are three important elements in kNN algorithm. A mathematical model is presented to describe kNN algorithm as follows:

$$\textit{Majority Voting: } y' = \textit{argmax}_v \sum_{(xi,yi)\in Dz} I(v = y_i) \qquad (2)$$

Input: Data set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$;

Base learning algorithm $\mathcal{L}$;

Number of learning rounds T.

Process:

$D_1(i) = 1/m.$ % Initialize the weight distribution

for $t = 1, \cdots, T$:

$h_t = \mathcal{L}(\mathcal{D}, D_t);$ % Train a weak learner h_t from $\mathcal{D}$ using distribution D_t

$\epsilon_t = \Pr_{i\sim D_t}[h_t(x_i) \neq y_i)];$ % Measure the error of h_t

$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right);$ % Determine the weight of h_t

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(x_i) = y_i \\ \exp(\alpha_t) & \text{if } h_t(x_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$ % Update the distribution, where Z_t is % a normalization factor which enables D_{t+1} be a distribution

end.

Output: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Figure. 1: AdaBoost pseudo-code [14]

where v is a class label, and yi is a class label of ith nearest neighbors. Meanwhile, I(.) is a function that yields 1 if true, and 0 is false [14].

kNN is a lazy learning; it only uses quick training phase. So, this algorithm is cheap in building the model and easy to understand. Besides its simplicity and robustness, kNN has some issues like determining k value and the approach to combining class labels.

2.5. Naive Bayes

Naive Bayes [23,24] is one of classic data mining algorithm. It is easy to construct and to involve simple iterative schemes. Naive Bayes may not be the best classification method, but it is well known for the robustness and accuracy [14]. The mathematical model of Naive Bayes is shown as follows.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (3)$$

where P(c|x) is the posterior probability of class prior probability P(c), and P(x|c) is the probability of predictor given class and P(x) is the prior probability of predictor.

2.6. CART

CART (Classification and Regression Trees) [25] is a part of decision tree method that was introduced in 1984. CART is a binary partitioning function that deals with continuous and nominal attributes. It processes the data in the raw form and trees are grown until the size is maximum without stopping the rule until it reaches stopping condition. Then, it prunes the branch which has the less impact to the tree. This mechanism yields invariant trees from any order which preserves transformation of attributes.

In practical, CART algorithms can be used in many sectors, but frequently mentioned in electrical engineering, financial, and biology research.

2.7.Random Forest

Random Forest is a member of decision tree algorithm family. It modifies the previous work of decision trees in constructing the classification trees. Random Forest splits nodes using the best among a subset of predictors which chosen randomly. The algorithm steps are as follows. First, draw the ntree bootstrap samples. Then, for each sample, grow an unpruned classification tree: at each node, randomly sample mtry of the predictors and choose the best split from those variables. Then, the prediction can be made by aggregating the prediction of ntree trees with majority votes [26].

From some research, Random Forest has some advantages like: can be used in multi-class problems, good predictive performance, does not overfit and can handle a mixture of categorical and continuous parameters. But somehow, Random Forest is not too frequently used in medical microarray studies [27].

2.8. Bagging Algorithm

Bagging algorithm is a method that combines the outputs of different models to make more reliable results [28]. This idea is depicted in Figure 2 below.

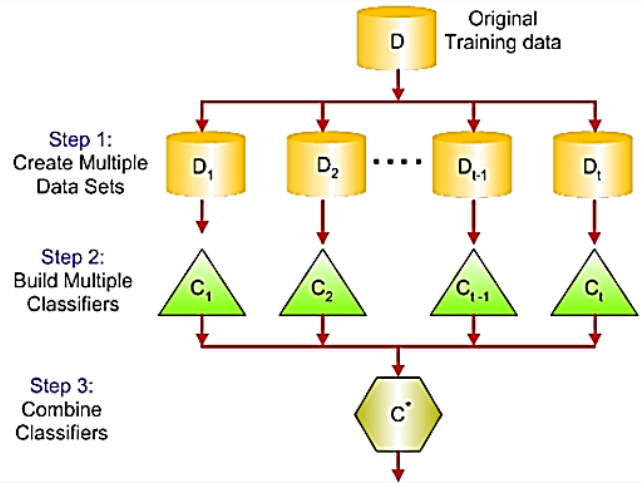

Figure. 2. Combining multiple models [28]

Bagging algorithm reduces the instability of combined methods with process simulation to the training data. The original dataset is modified by removing some data and duplicating the others. As for the substitution, data are selected randomly to create the new ones with the same size. The process of replicating and removing data is a part of how Bagging algorithm works. Until now, Bagging is known as an effective algorithm for many purposes, including medical diagnosis [11].

2.9. Logistic Regression

Logistic regression is a model that classifies training data into two conditions like "0" and "1" refers to Bernoulli distribution [29]. When the expected output is more than two outcomes, multinomial logistic regression is chosen.

Logistic regression is used in many sectors of research, from marketing applications, social sciences, and healthcare research. For example, a research [30] uses logistic regression to analyze congenital heart disease. The other example is [31] when logistic regression is used to predict whether a customer will purchase or terminate a product delivered by a company.

2.10. MLP

MLP or feed-forward artificial neural network consists of some neurons which connected with connecting nodes (Figure 3). Those neurons are arranged within layers: one input layers, one or more hidden layer, and one output layer. The input layer receives a signal from outside, passes it through hidden layers, and finally reaches the output layer.

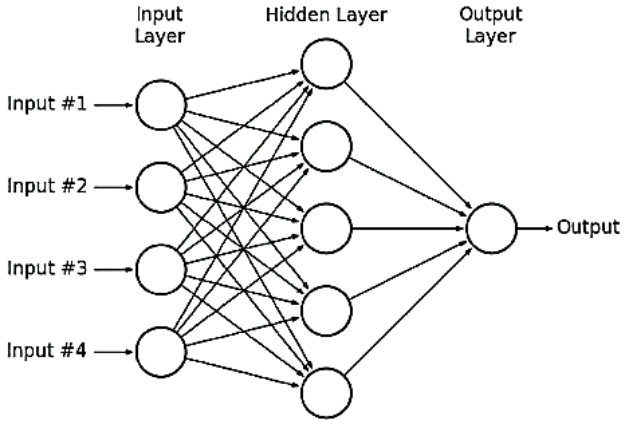

Figure 3. A Multilayer Perceptron [32]

For training purpose, MLP uses the backpropagation algorithm, which works as summarized procedure as follows [32]:

1. Set the network weights.
2. Determine the first input vector (from the dataset) to the network.
3. Deploy the input through the network, to get an output.
4. Check the actual output and compare with expected one, to get the error signal.
5. Error signal should be propagated back through the network.
6. To reduce error, weights should be adjusted.
7. Repeat the procedure from step 2 using next input, to get the optimal result.

MLP is used in many sectors as well, in general for function approximation, prediction, and pattern classification. It is shown that this method is superior to any traditional classification methods [33].

3. METHODS AND PROCEDURES

3.1 Dataset

In this study, the data for training set were collected from Harapan Kita Hospital (HKH) Jakarta, the biggest cardiovascular hospital in Indonesia in 2016 period. As many as 450 records are collected, while the parameters of the data used in this research are written in Table 1.

For the analysis purposes, the text data format should be converted to binary, integer, or code so the data can be analyzed. The cardiovascular specialist's expertise is needed in this work.

First, Symptom parameter should be converted to binary data. There are main symptoms that related to cardiovascular disease which are: chest pain (CP), shortness of breath (SB), palpitations (PA), and cold sweat (CS) [34]. So, the text content of Symptom parameter

Table 1. Data parameters and formats from HKH

No	Parameter	Format
1	Sex	Binary (Male or female)
2	Age	Integer (Years old)
3	Symptom & Additional symptom	Text
4	Blood pressure	Integer (Systolic & diastolic, mmHg)
5	Heart rate	Integer (beat per minute)
6	EKG	Text
7	Diagnosis	Text

can be converted to binary data whether a patient experience CP (Yes/No), SB (Yes/No), PA (Yes/No), and CS (Yes/No).

Next, for EKG text data are needed to be converted as well. According to [35], EKG data can be categorized into 8 important sub-parameters which are: EKG Rhythm (ER), EKG Axis (EA), EKG P-Wave (EPW), EKG PR Interval (EPR), EKG QRS Duration (EQ), EKG Bundle Branch Block (EBB), EKG ST-T Changes (EST), and EKG T-Wave (ETW).

The last conversion process needed of HKH dataset is for Diagnosis parameter as an output which has multi-class labels. The data should be changed to ICD-10 coding. ICD-10 is a reference of disease coding which has been approved by World Health Organization [36]. Finally, 16 parameters (15 input and 1 output) are produced from the conversion process, as written in Table 2.

Table 2. Conversion result of HKH parameter formats

No	Parameter	Format
1	Sex	Binary (Male or female)
2	Age	Integer (Years old)
3	CP	Binary (Yes/No)
4	SB	Binary (Yes/No)
5	PA	Binary (Yes/No)
6	CS	Binary (Yes/No)
7	Blood pressure	Integer (Systolic & diastolic, mmHg)
8	Heart rate	Integer (beat per minute)
9	EA	Integer
10	EPW	Integer
11	EPR	Integer
12	EQ	Integer
13	EBB	Integer
14	EST	Integer
15	ETW	Integer
16	Diagnosis	Code (ICD-10)

3.2. Data Mining Analysis

For data mining analysis in this study, a tool named WEKA is used. WEKA is a popular tool for data mining analysis, and it is widely used since it is a free license software [37]. Ten of data mining algorithms mentioned in this paper will be tested with WEKA. In general, the 10-fold cross-validation option is used for data training/testing, and classifier options are left with default settings. The methodologies used in WEKA for each algorithm are summarized in Table 3.

Table 3. Summary of WEKA methods for analysis

No	Analysis	Classifier Options
1	C4.5	• confidenceFactor=0.25 • minNumObj=2
2	SVM	• SVMType=C-SVC • Degree=3 • Gamma=0 • Epsilon=0.001 • KernelType=radial basis function
3	AdaBoost	• Classifier=DecisionStump • numIterations=10 • seed=1 • weightThreshold=100
4	kNN	• k=1 • distanceWeighting=No • NNSearchAlgorithm=LinearNNSearch
5	Naive Bayes	• Cross-validation folds=10
6	CART	• Classifier=SimpleCart • Seed=1 • minNumObj=2 • numFoldsPruning=5
7	Random Forest	• numTrees=100 • maxDepth=0 • seed=1
8	Bagging Algorithm	• bagSizePercent=100 • classifier=REPTree • numIterations=10 • seed=1
9	Logistic Regression	• maxIts=1 • ridge=1.0E-8
10	MLP	• learningRate=0.3 • momentum=0.2 • seed=0 • trainingTime=500

4. RESULTS

4.1 Prediction Accuracy

After performing WEKA analysis for ten selected data mining algorithms, the result of each algorithms' accuracy is summarized in Table 4.

Table 4. Result of algorithm's prediction accuracy

Rank	Algorithm	Accuracy
1	Random Forest	78.0%
2	kNN	71.6%
3	MLP	63.8%
4	Bagging Algorithm	63.1%
5	C4.5	62.9%
6	Logistic	62.4%
7	CART	60.5%
8	Naive Bayes	50.4%
9	AdaBoost	46%
10	SVM	45.1%

From the result, Random Forest comes with the best accuracy of 78.0%, followed by kNN (71.6%), MLP (63.8%), Bagging Algorithm (63.1%), C4.5 (62.9%), Logistic Regression (62.4%),

CART (60.5%), Naive Bayes (50.4%), AdaBoost (46%), and the last is SVM (45.1%).

4.2. Processing Speed

Another important key performance index of an algorithm is its speed. Most applications now are requiring a fast response or even real-time. The speed of ten algorithms here is measured and summarized in Table 5. The time recorded here is using a personal computer with Intel Core i5 @1.6 GHz CPU and 12GB RAM for WEKA simulation.

Table 5. Result of algorithm's speed

Rank	Algorithm	Analysis duration
1	AdaBoost	< 1 s
	kNN	< 1 s
	Naive Bayes	< 1 s
2	Bagging Algorithm	1 s
	C4.5	1 s
	CART	1 s
7	Random Forest	8 s
8	MLP	2m 57s
9	Logistic	3m 7s
10	SVM	4m 41s

From speed measurement, AdaBoost, kNN, and Naive Bayes are on top of the list with less than 1 second processing time. Bagging algorithm, C4.5, and CART spent 1 second. Meanwhile, the rest of the list are: Random Forest (6 seconds), MLP (2 minutes 29 seconds), Logistic Regression (3 minutes 9 seconds), and the last is SVM (4 minutes 16 seconds).

4.3. Discussion

From the experiments using HKH data, Random Forest is the best algorithm regarding accuracy (78.0%) but failed to perform very well regarding speed (8 seconds). In this study, the computer used for simulation has a good specification for today's standard. If a lower specification one is used, the processing time can be longer. Meanwhile, kNN shows more attractive results. Although the accuracy is only the second best (71.6%), the processing time is less than 1 second, or almost instantaneous. MLP is not quite good in accuracy (63.8%) and the speed is disappointing (2 minutes 57 seconds). Bagging algorithm and C4.5 share a similar performance with about 63% accuracy and 1 second processing time. Logistic regression gives 62.4% and 3 minutes 7 seconds processing time. CART has a low accuracy (60.5%) but the speed is very good (1 second). Naive Bayes has a great processing time under 1 second, but the accuracy is only 50.4%. Similar with AdaBoost who has almost instantaneous processing time (less than 1 second) but only 46% of accuracy. The last, SVM shows the worst result both in accuracy (45.1%) and speed (4 minutes 41 seconds).

These results gave a description on how each algorithm performs in handling a multi-class expected output. kNN provides us attractive overall performance with a fair accuracy and fast speed. Random Forest is the best regarding prediction, although the speed is not too good.

5. CONCLUSION

Data mining method now has been widely used in many sectors, including healthcare. This paper aims to check the performance of top ten data mining algorithm (C4.5, SVM, AdaBoost, kNN, Naive Bayes, CART, Random Forest, Bagging Algorithm, Logistic Regression, and MLP) with a training dataset collected (450 records) from Harapan Kita Hospital in Jakarta. The dataset has 15 input parameters and one multi-class output parameter after pre-processed steps. The tool used for data mining analysis is WEKA. The key performance indexes of these algorithms are the accuracy and speed.

The highlights of this experiment result are as follows. For the prediction accuracy, the top three algorithms are Random Forest (78.0%), kNN (71.6%), and MLP (63.8%). While the top three regarding of speed are AdaBoost, kNN, and Naive Bayes (all under 1 second). We can say that kNN gave us attractive result since it is always mentioned in the top three performance for accuracy and speed.

6. REFERENCES

[1] Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515-524.

[2] Liao, S. H., Chu, P. H., Chen, Y. J., & Chang, C. C. (2012). Mining customer knowledge for exploring online group buying behavior. Expert Systems with Applications, 39(3), 3708-3716.

[3] Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. Expert Systems with Applications, 39(12), 11243-11249.

[4] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569.

[5] Rajesh, K., and Sangeetha, V. "Application of data mining methods and techniques for diabetes diagnosis." International Journal of Engineering and Innovative Technology (IJEIT) 2.3 (2012).

[6] Sung, S. F., Hsieh, C. Y., Yang, Y. H. K., Lin, H. J., Chen, C. H., Chen, Y. W., & Hu, Y. H. (2015). Developing a stroke severity index based on administrative data was feasible using data mining techniques. Journal of clinical epidemiology, 68(11), 1292-1300.

[7] Kharya, Shweta. "Using data mining techniques for diagnosis and prognosis of cancer disease." arXiv preprint arXiv:1205.1923 (2012).

[8] European Public Health Alliance (EPHA). What are the leading causes of death in the EU? Accessed via: www.epha.org/a/235.2 (2014).

[9] Mann, D. L., Zipes, D. P., Libby, P., & Bonow, R. O. Braunwald's heart disease: a textbook of cardiovascular medicine. Elsevier Health Sciences, (pp 2) (2014).

[10] Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B. & Turner, M. B. Executive summary: heart disease and stroke statistics—2012 update, a report from the American Heart Association. Circulation, 125(1), 188-197 (2012).

[11] Tu, My Chau, Dongil Shin, and Dongkyoo Shin. "Effective diagnosis of heart disease through bagging approach." Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on. IEEE, 2009.

[12] Kumari, Milan, and Sunila Godara. "Comparative study of data mining classification methods in cardiovascular disease prediction 1." (2011).

[13] Enriko, I., Wibisono, G., & Gunawan, D. Designing machine-to-machine (M2M) system in health-cure modeling for

cardiovascular disease patients: Initial study. In Information and Communication Technology (ICoICT), 2015 3rd International Conference (pp. 528-532). IEEE (2015).

[14] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Steinberg, D. Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37 (2008).

[15] Quinlan, J. Ross. C4. 5: programs for machine learning. Elsevier, 2014.

[16] Amiri, B., Hossain, L., & Crawford, J.. "A hybrid evolutionary algorithm based on HSA and CLS for multi-objective community detection in complex networks." Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. IEEE, 2012.

[17] Slocum, Mary. "Decision making using ID3 algorithm." Insight: River Academic J 8.2 (2012).

[18] Lu, Y., Boukharouba, K., Boonært, J., Fleury, A., & Lecoeuche, S. (2014). Application of an incremental SVM algorithm for on-line human recognition from video surveillance using texture and color features. Neurocomputing, 126, 132-140.

[19] Ying, C., Qi-Guang, M. I. A. O., Jia-Chen, L. I. U., & Lin, G. A. O. (2013). Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica, 39(6), 745-758.Dietterich TG (1997) Machine learning: Four current directions. AI Mag 18(4):97–136

[20] Polikar, R. (2012). Ensemble learning. In Ensemble machine learning (pp. 1-34). Springer US.

[21] Larose, D.T. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.

[22] Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. Journal of advanced research, 4(4), 331-344.

[23] Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 194-201). Springer Berlin Heidelberg.

[24] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications, 6(2), 256-261.

[25] Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). The CART decision tree for mining data streams. Information Sciences, 266, 1-15.

[26] Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493-507.

[27] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 3.

[28] H.W. Ian, E. Frank. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[29] Freedman, David A. Statistical models: theory and practice. Cambridge University Press (p128), 2009.

[30] Aydin, S. I., Seiden, H. S., Blaufox, A. D., Parnell, V. A., Choudhury, T., Punnoose, A., & Schneider, J. (2012). Acute kidney injury after surgery for congenital heart disease. The Annals of thoracic surgery, 94(5), 1589-1595.

[31] Lee, W. I., Chen, C. W., Chen, K. H., Chen, T. H., & Liu, C. C. (2012). Comparative study on the forecast of fresh food sales using logistic regression, moving average and BPNN methods. Journal of Marine Science and Technology, 20(2), 142-152.

[32] Mohamed, H., Negm, A., Zahran, M., & Saavedra, O. C. (2015). Assessment of artificial neural network for bathymetry estimation using High Resolution Satellite imagery in Shallow Lakes: case study El Burullus Lake. In International water technology conference (pp. 12-14).

[33] Zare, M., Pourghasemi, H. R., Vafakhah, M., & Pradhan, B. (2013). Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. Arabian Journal of Geosciences, 6(8), 2873-2888.

[34] Irmalita JD, Andrianto SB, Tobing DPL, Firman D, Firdaus I. Pedoman tatalaksana sindrom koroner akut, 2015.

[35] Gertsch M. The Normal ECG and its (Normal) variants. InThe ECG 2004 (pp. 19-43). Springer Berlin Heidelberg.

[36] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care. 2005 Nov 1:1130-9.

[37] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.