Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters

I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan
Dept. of Electrical Engineering, Universitas Indonesia, Indonesia.
i.ketut42@ui.ac.id

***Abstract*—Heart disease is the primary cause of death nowadays. Treatments of heart disease patients have been advanced, for example with machine-to-machine (M2M) technology to enable remote patient monitoring. To use M2M to take care remote heart disease patient, his/her medical condition should be measured periodically at home. Thus, it is difficult to perform complex tests which need physicians to help. Meanwhile, heart disease can be predicted by analysing some of patient's health parameters. With help of data mining techniques, heart disease prediction can be improved. There are some algorithms that have been used for this purpose like Naive Bayes, Decision Tree, and k-Nearest Neighbor (KNN). This study aims to use data mining techniques in heart disease prediction, with simplifying parameters to be used, so they can be used in M2M remote patient monitoring purpose. KNN is used with parameter weighting method to improve accuracy. Only 8 parameters are used (out of 13 parameters recommended), since they are simple and instant parameters that can be measured at home. The result shows that the accuracy of these 8 parameters using KNN algorithm are good enough, comparing to 13 parameters with KNN, or even other algorithms like Naive Bayes and Decision Tree.**

***Index Terms*—Heart Disease Prediction; k-Nearest Neighbor; Data Mining; Machine to Machine.**

I. INTRODUCTION

Heart disease is the primary cause of death of humankind nowadays. It is reported in USA that cardiovascular death is about one-third of overall death [1]. Another study stated that in Europe the percentage of death caused by cardiovascular disease (CVD) is 35% [2]. The condition is similar in low-and-middle-income countries, where about 28% of mortality cause is CVD [3].

In emergent nations, the quality of healthcare services still needs to be improved. Like in Indonesia, we are lacking of medical practitioners where the ratio is 0.36 doctor per 1000 residents [4]. Thus, many research and innovations in healthcare service improvements are thriving, for example: the use of machine-to-machine (M2M) technology in patient monitoring [5-9]. With its intensive developments, M2M technology will be massively used in various fields, including healthcare.

While many research have been done in medical discipline related to CVD, data mining techniques have been used in healthcare diagnosis as well. Data mining is a process of extracting or exploring large size of data to gain knowledge, pattern, or relationship which involves statistical analysis, machine learning, and database management [10,11].

Some research in data mining for CVD are mostly related to heart disease predictions, for example [12] which compared three data mining algorithms: CART, ID3, and Decision Table to predict whether a patient will have heart disease or not. A dataset from California University, Irvine (UCI) is taken to do the analysis, using 10 out of 76 parameters in available. The highlighted results were the accuracy of the algorithms: CART = 83.5%, ID3 = 72.9% and Decision Table = 82.5%. Other related study is [13]. The study used database software called MongoDB with Naive Bayes, Decision List, and KNN algorithm to predict patient's heart disease. They also used sample dataset from UCI with 13 ouf of 76 parameters available. The result is the accuracy of algorithms used: Naive Bayes 52.33%, Decision List 52%, and KNN gives 45.67. The last example is [14] which aimed to predict heart disease with K-means clustering and MAFIA algorithm. They use UCI dataset with 11 parameters chosen. They produced important results: 74% to 89% accuracy with different techniques proposed.

This study proposes heart disease prediction using KNN with instant measurement parameters. KNN is one of the top data mining algorithm [15,16] which frequently used in disease prediction method [17, 18]. The parameters are body vital signs that can be measured instantly where, for example, any invasive procedures, fasting, or complex procedures (like MRI or X-ray scanning) are not needed. The benefit of instant parameters is they can be acquired through sensors for patients who are treated at home, if we want to implement M2M in helping CVD patients.

II. LITERATURE REVIEW

These years, data mining has become popular in many fields of industry, thanks to its purpose to convert large to become valuable information [10]. Examples of data mining use are mentioned below:

- A retail store arranges the merchandise by seeing its customer buying behaviours and products connection information [19].
- Analysis of churn pattern in telecommunication business competition [20].
- Analysis of web browsing pattern to optimize a website

design [21].
- Analysis for financial fraud detection [22].
- Analysis for a specific disease diagnosis [23].

There are many popular data mining algorithms, especially classification techniques, which each of them has superiority and weakness as well, three of them are: Naive Bayes, Decision Tree, and k-Nearest Neighbor (KNN) [15]. Naive Bayes is a simple, robust, and well-performed classification technique [15, 24]. Basically, it is based on Bayes' theorem to calculate posterior probability P(c|x) from class prior probability P(c), probability of predictor given class P(x|c) and the prior probability of predictor [15, 24, 25], or:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Naive Bayes model is widely used in areas like spam filtering, text classification, even medical diagnosis. It gets much attention amongst statistics experts which resulted in algorithm modifications [15, 24].

Meanwhile, Decision Tree algorithms are one of the top in data mining world as well, thanks to its speed in training phase and clear modeling. Decision Tree works by classifying trained data to form a tree. This tree is formed in training phase to see how accurate the classifier for test data. Then the test data will be classified using the tree [26]. Some more specific techniques of Decision Tree algorithms have been invented by data mining experts. The prominent ones are:

1. CART (Classification And Regression Tree)
2. ID3 (Iterative Dichotomiser 3)
3. C4.5 (development of ID3)
4. Random Forest

The other data mining mentioned here is k-Nearest Neighbor (KNN). KNN is a basic and simple classification technique which frequently used in many studies, especially when there is only few or no information about the data distribution [27]. It is a non parametric algorithm, means that KNN does not make presumptions about distribution of data used in analysis. It fits in practical environments, where oftentimes real data do not follow theoritical statistics like normal distribution. KNN also called a lazy algorithm, or it only uses quick training phase. KNN does not make generalization which implies that KNN maintains all training data.

Euclidean distance is usually used in KNN classifier to calculate the similarity between training and test data. It is calculated with formula below [28]:

$$d_2(X,Y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2} \quad (2)$$

There are many researches about using data mining for medical purpose, especially in heart disease, for example Chaurasia's research: "Early Prediction of Heart Diseases Using Data Mining Techniques " [12]. He compared 3 data mining algorithms: CART, ID3, and Decision Table to predict whether a patient will have heart disease or not. A dataset from California University, Irvine (UCI) is taken to do the analysis, it used 10 out of 76 parameters in this study, they are: (1) Age, (2) Sex, (3) CP, (4) Trestbps, (5) Chol, (6) Fbs, (7) Restecg, (8) Thalach, (9) Exang, and (10) Slope. The highlighted results were the accuracy of the algorithms: CART = 83.5%, ID3 = 72.9% and DT = 82.5%.

Other related study is Jarad et al: "Intelligent Heart Disease Prediction System With MongoDB" [13]. They used database software called MongoDB with Naive Bayes, Decision List, and KNN algorithm to predict patient's heart disease. They also used sample dataset from UCI with 13 ouf of 76 parameters available: (1) Age, (2) Sex, (3) CP, (4) Trestbps, (5) Chol, (6) Fbs, (7) Restecg, (8) Thalach, (9) Exang, (10) Oldpeak, (11) Slope, (12) Ca, and (13) Thal. The result of this study gives accuracy of algorithms used: Naive Bayes 52.33%, Decision List 52%, and KNN gives 45.67 accuracy.

The last example is a study done by Karthiga et al: "Heart Disease Analysis System Using Data Mining Techniques" [14]. They work was to predict heart disease as well, with K-means clustering and MAFIA algorithm. They use UCI dataset with 11 parameters chosen: (1) Age, (2) Sex, (3) Slope, (4) Famhist, (5) Fbs, (6) Painloc, (7) Thal, (8) Chol, (9) Trestbps, (10) Exang, and (11) Thalach. They produced important results: 74% to 89% accuracy with different techniques proposed.

III. MATERIALS AND METHODOLOGY

A. Dataset

In this research we use a dataset from UCI [29] called Hungarian dataset, which has most data records (293 records after removing incomplete data). There are totally 76 parameters in the dataset but we only use 8 as written in Table 1.

Table 1
Parameters from UCI Dataset Used in This Study

No	Parameter	Description
1	Age	Age of the patient, in year
2	Sex	0 = Female, 1 = Male
3	CP	Chest Pain type: 1 = Typical angina 2 = Atypical angina 3 = Non-angina pain 4 = Asymptomatic
4	Trestbps	Resting blood pressure systolic
5	Trestbpd	Resting blood pressure diastolic
6	Restecg	Resting ECG: 0 = Normal 1 = Having ST-T wave abnormality 2 = Showing probable or definite left ventricular hypertrophy by Estes' criteria
7	Thalrest	Resting heart rate
8	Exang	Exercise induced angina: 0 = No; 1 = Yes

The final parameter is diagnosis result which is the prediction result, whether a patient is healthy (0) or have heart disease (1).

The reason to choose those 8 parameters is that they can be measured instantly, means:

1. Patient is not required to take certain procedures like fasting.
2. The measurement is not invasive, only external body measurements included which can be done by simple medical devices.
3. Some parameters are not acquired by medical device

(age, sex, CP and exang), so there should be an interview to get the data. This can be done through a communication device like smartphone with a special application.

Other important reason to choose those parameters is that the real data in hospital frequently incomplete. It can be understood since patient with heart attack sometimes need quick help from paramedics then they ignore to fill in the data form completely. For this purpose, we have conducted a survey in Harapan Kita Heart & Cardiovascular Hospital (HARKIT), Jakarta. We have collected 387 medical records with the format described in Table 2.

Table 2
Medical Record Format in HARKIT

No	Field	Description	Complete / Not
1	Medical record ID	Patient's medical record ID	Complete (387 records)
2	Sex	Sex	Complete (387 records)
3	Age	Age in years	Complete (387 records)
4	Symptom	Patient's complaint description (pain, illness, etc)	Complete (387 records)
5	Additional Symptom	Additional patient's complaints	Almost complete (351 records)
6	Blood pressure	Blood pressure sys & dia	Almost complete (381 records)
7	Heart rate	Patient's heart rate	Almost complete (381 records)
8	Cholesterol level	Cholesterol level	Not Complete (15 records)
9	Trop T	Troponin T rate	Not Complete (100 records)
10	CKMB	Creatine Kinase MB level	Not Complete (74 records)
11	GDP	GDP level	Not Complete (149 records)
12	Echo	Echocardiogram test result	Not Complete (84 records)

If we check the data format from HARKIT, then we see that some parameters are identical with dataset fields from UCI which are: Sex, Age, Blood Pressure (Trestbps and Trestbpd in UCI data), Heart Rate (Thalrest), and EKG (Restecg). While, CP and Exang (in UCI data) can be inferred from Symptom and Additional Symptom in HARKIT data. We ignore incomplete ones and other data in HARKIT data that cannot be provided by UCI data.

B. Data Mining Analysis

In this research we mainly use Microsoft Excel / Macro Visual Basic (Excel Macro) for doing the KNN analysis. To support the main analysis, we use WEKA, a data mining tool popularly used in classification techniques [30], version 3.6.12.

The procedures in doing KNN with Excel Macro are described below.

1. We start the analysis by dividing the data we have into training data and test data. In each analysis, 90% of the data are for training data and 10% are for test data. We do the analysis 10 times so we have all data act as test data once. In this case, we have 293 data so, first analysis we have record number 1 to 29 as test data, and record number 30 to 293 as training data. Then in second analysis, we have record number 30 to 58 as test data, while record number 1 to 29 and 59 to 293 as training data, and so on. This mechanism is known as 10-folds cross validation, as depicted in Figure 1 below. In the end, we have 10 results from 10-folds cross validation. The final result is average of them.

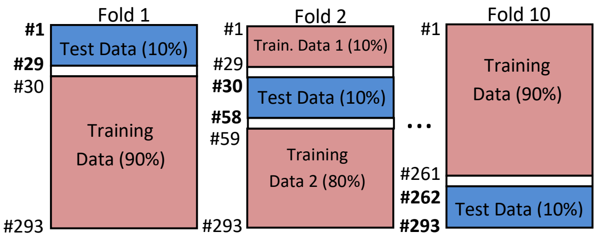

Figure 1: 10-Folds Cross Validation Mechanism

2. In each fold, we calculate the similarity score of each training data compared to test data. If n = number of test data (in this case is 10% of 292 or about 29), m = number of training data (in this case is 90% of 292 or about 263), and a = number of parameters (in this case is 8). The formula of calculating similarity score for this case is:

   Score of each parameter =

   $$\sum_{j=1}^{n} \sum_{i=1}^{m} \frac{|\,Test(j) - Training(i)\,|}{max(m) - min(m)} \, W$$

   Score for a test data = Score parameter 1 + Score parameter 2 + ... + Score parameter 8

   For example if a test value equals to a training value to be compared, the score will be 0 (matched). In contrast, score = 1 means that test value and training value are the farthest each other.

3. After all test data (in one fold) is determined, we should sort them so we know which training datum is the closest (or most similar) to the test datum. One closest datum will be considered as the most similar for k = 1, three closest data are the most similar ones for k = 3, and so on. Let's see an example below:
   For test data 1 (which heart disease prediction is "N"):
   - For k = 1: the 1st closest datum is training data #33, score = 0.100 (which heart disease prediction = "N").
     So the training datum prediction is accurate since its heart disease prediction ("N") is the same with the test data.
   - For k = 3:
     The 1st closest datum is training data #33,
     score = 0.100 (heart disease prediction = "N");
     The 2nd closest datum is training data #65,
     score = 0.111 (heart disease prediction = "Y");
     The 3rd closest datum is training data #78,

score = 0.114 (heart disease prediction = "Y").
Since there are three data and they give different predictions, we should vote them and it yields heart prediction = "Y", which is not accurate because the test data prediction is "N".

- And so on.

Do this until the last test data in a fold.

4. For each fold (29 test data), we could determine the prediction accuracy. For example if 20 data are accurate and 9 data are inaccurate, then the accuracy is 69%.
5. Do step 2 to 4 for all 10 folds.

The flowchart of these steps is depicted in Figure 2 below.

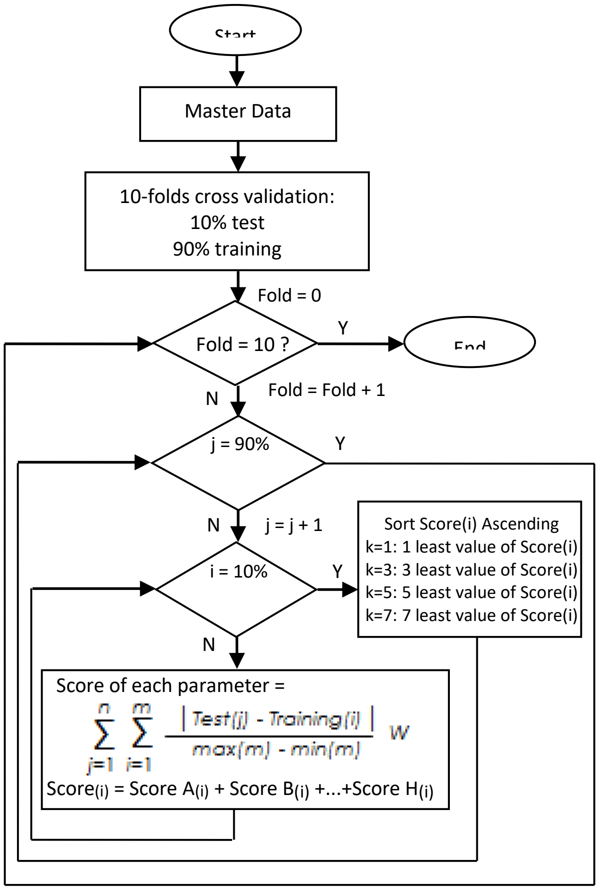
Start
Master Data
10-folds cross validation:
10% test
90% training
Fold = 0
Fold = 10 ?
Y
End
N
Fold = Fold + 1
j = 90%
Y
N
j = j + 1
i = 10%
Y
N
Sort Score(i) Ascending
k=1: 1 least value of Score(i)
k=3: 3 least value of Score(i)
k=5: 5 least value of Score(i)
k=7: 7 least value of Score(i)
Score of each parameter =
$\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{|Test(j) - Training(i)|}{max(m) - min(m)} W$
Score(i) = Score A(i) + Score B(i) +...+Score H(i)

Figure 2: Flowchart of KNN Calculation Using Excel Macro

With KNN algorithm, we have chance to change the parameter's weight. It means that, we may assume that some parameters are more important or making more impact than others. Among 8 parameters we use, we can categorize them our data into 2 categories, one is "non-medical" parameters (Age and Sex) and the other is "medical" parameters (CP, Trestbps, Trestbpd, Restecg, Thalrest, and Exang). We may think that medical parameters are more important than non-medical, which we will see in our experimental result. In this research we perform some calculations for different parameter weighting to see which one gives the best result. Along with weighting, we should find the value of "k" so it gives the best classification result. Since it is a 2-choice classification ("yes" or "no" prediction), the "k" value will be an odd number (1,3,5, and so on).

To complete the analysis, we compare the result of KNN analysis with other data mining algorithms such as Naive Bayes and Decision Tree. Since there are many different algorithms in Decision Tree, we use the popular one: Simple CART. WEKA software helps to do the analysis since it has a great number of library for data mining algorithms. Summary of methodology used in this study is written in Table 3 below.

Table 3
Summary of Methodology in This Research

Analysis	Tools	Remarks
KNN	Excel Macro	• 10-folds validation method is used • k values for testing: k=1,3,5, and 7 • Different parameter weightings are calculated
Naive Bayes	WEKA	Test option: 10-folds cross validation method is used
Decision Tree (Simple CART)	WEKA	Classifier options: • Debug = false • Heuristic = true • minNumObj = 2.0 • NumFoldsPruning = 5 • seed = 1 • sizePer = 1.0 • useOneSE = False • usePrune = True

IV. RESULTS AND DISCUSSION

A. *8-Parameters KNN Experiments*

Before we perform the KNN analysis and compare with other method, we may take a look at the parameters/variables. We want to know how important each variable is, in order to know which variables are more important than others. This information could be a reference when we want to do the parameter weighting in KNN. There are many ways to compute the variable importance, one of them is with Chi-Square attribute evaluation. With WEKA tool, we can determine the attribute importance with Chi-Square attribute evaluation with results informed in Table 4.

Table 4
Results of 8 Parameters Variable Importance Test Using WEKA

Rank	Score	Attribute
1	100.456	8 exang
2	90.583	3 cp
3	21.876	2 sex
4	0	7 thalrest
5	0	4 trestbps
6	0	5 trestbpd
7	0	6 restecg
8	0	1 age

From the test now we know that 3 most important variables are: Exang, CP, and Sex, while other variables are concluded as less important. Later in KNN weighting experiments we will give those 3 variables with weight = 2 and others = 1. Based from this result, later we will test KNN algorithm with more weight on "Top 2 Parameters" (Exang and CP) and "Top 3 Parameters" (Exang, CP, and Sex).

One more important thing in KNN algorithm is how to determine the optimum k parameter. First, k should be an odd number since we have to vote the nearest neighbors into two classes (YES or NO) so if we choose even numbers the result can be tied [31]. Second, many research have been done to determine optimum k parameter but there is no ultimate method to determine "optimum k parameters" so one method that can be used is to select k parameter using k=1 until k = square root of training data (k=1, k=3, k=7, ... k= √n) [32]. In this experiment, since the dataset consists of 293 records, and we used 10-folds cross-validation, it means that 90% (or 264 records) are training data and 10% (or 29 records) are test data. Thus, maximum number of k parameter is square root of 264 which is 16.25. Finally we can determine that the maximum number of k parameter is 15, for this experiment, as written in Table 5.

Table 5
Accuracy Table for KNN Algorithm Using 8 Parameters with Parameter Weightings Based on Variable Importance Test

k	Without parameter weightings	Chi-Squared Top 3 Parameters (12211112)	Chi-Squared Top 2 Parameters (11211112)
k=15	80.82%	81.16%	80.82%
k=13	81.16%	79.45%	80.82%
k=11	80.82%	80.14%	**81.85%**
k=9	80.48%	79.11%	80.82%
k=7	80.48%	78.77%	80.14%
k=5	79.79%	78.77%	79.11%
k=3	78.42%	75.68%	75.68%
k=1	74.32%	72.26%	72.26%

From the table we can see that the best result of the experiments is 81.85% for Chi-Squared Top 2 Variables with k=11.

B. 13 Parameters KNN Experiments

We do the same steps with 8 parameters, now using 13 parameters. For these 13 parameters, the weighting sequence is (1) Age, (2) Sex, (3) CP, (4) Trestbps, (5) Chol, (6) FBS, (7) Restecg, (8) Thalach, (9) Exang, (10) Oldpeak, (11) Slope, (12) CA, and (13) Thal.

First we conduct the Chi Squared attribute evaluation using WEKA tools to determine which variables are more important than others, that mentioned in Table 6.

Table 6
Result of 13 Parameters Variable Importance Test Using WEKA

Rank	Score	Attribute
1	110.334	11 slope
2	100.456	9 exang
3	90.583	3 cp
4	90.227	10 oldpeak
5	29.239	8 thalach
6	21.876	2 sex
7	0	4 trestbps
8	0	13 thal
9	0	7 restecg
10	0	5 chol
11	0	6 fbs
12	0	12 ca
13	0	1 age

From the test we found that top 6 variables are: Slope, Exang, CP, Oldpeak, Thalach, and Sex. We can state that these 6 variables are more important than 7 others, so in KNN weighting experiments we will give the top 6 variables weight = 2 while 7 others are = 1. Meanwhile, among those 6 top variables there are 4 variables which have more than 90 points: Slope, Exang, CP, and Oldpeak. Then we may do the experiment with giving those top 4 variables with weight = 2, while 9 others with weight = 1.

Then we do the KNN weighting experiments to check the accuracy. The results are concluded in Table 7.

Table 7
Accuracy Table for KNN Algorithm Using 13 Parameters with Parameter Weightings Based on Variable Importance Test

k	Without parameter weightings	Chi-Squared Top 6 Parameters (1221111222211)	Chi-Squared Top 4 Parameters (1121111122211)
k=15	79.93%	79.59%	79.93%
k=13	79.59%	79.25%	79.25%
k=11	78.57%	79.25%	78.23%
k=9	79.93%	78.91%	**80.61%**
k=7	78.23%	79.25%	78.23%
k=5	76.87%	79.93%	79.25%
k=3	78.57%	79.25%	78.23%
k=1	76.19%	77.55%	75.51%

From the table we can see that the best result of the experiments is 80.61% for Chi-Squared Top 4 Variables with k=9.

C. Naive Bayes and Decision Tree Experiments

After doing analysis with KNN method, we want check the results with Naive Bayes and Decision Tree algorithm. We perform the analysis with WEKA tool, for both 8 and 13 parameters. All results of Naive Bayes and Simple CART experiments are summarized in Table 8 below.

Table 8
The Accuracy Table of Naive Bayes and Simple CART Experiments for 8 and 13 Parameters

	Naive Bayes	Simple CART
8 parameters	74.49%	80.27%
13 parameters	79.93%	79.93%

D. Discussion

From all experiments performed with KNN (with 8 and 13 parameters), Naive Bayes and Decision Tree algorithms, we see that the accuracy results are not too far differed as seen in Table 9 below. In this research, the 8 parameters KNN gives the best result with 81.85% accuracy.

Table 9
Summary of KNN, Naive Bayes, and Simple CART Experiments for 8 and 13 Parameters

	KNN (Best Result)	Naive Bayes	Simple CART
8 parameters	81.85%	74.49%	80.27%
13 parameters	80.61%	79.93%	79.93%

We wrap up the discussion with a comparison of this study result with previous related studies [12,13,14]. In Table 10, we can see that this study is within the top results with 81.9% accuracy, while the best accuracy is 89% and the worst is 45.7%.

Table 10
Comparison of This Study Results with Previous Studies'

Study	Number of parameter	Method	Accuracy
[12]	10	CART	83.5%
		ID3	72.9%
		Decision Table	82.5%
[13]	13	Naive Bayes	52.3%
		Decision List	52.0%
		KNN	45.7%
[14]	11	K-means/MAFIA	74.6%
		K-means/MAFIA with ID3	83.0%
		K-means/MAFIA with ID3 & C4.5	89.0%
This study	8	KNN with parameter weighting	81.9%

V. CONCLUSION

Data mining technics have been used in many fields, one of them is healthcare. This paper's objective is to check whether heart attack prediction can be based on fewer parameters than what recommended on previous studies. We use 8 parameters (out of 13 recommended), which are: (1) Age, (2) Sex, (3) Chest pain, (4) Resting blood pressure systolic, (5) Resting blood pressure diastolic, (6) Resting ECG, (7) Resting heart rate, and (8) Exercise induced angina. The reasons to choose those parameters for this study are: they are simple measurements and consistently recorded in Harapan Kita Hospital, the biggest cardiovascular hospital in Indonesia.

Experiments using 8 parameters with KNN shows good accuracy if we compared with 13 parameters, even with other data mining algorithms like Naive Bayes and Decision Tree (in this research we use Simple CART). The benefit as the result from this study is: we can proof that 8 simple parameters are good enough to be used in heart attack prediction.

In our future research, it can be used as parameters in remote patient monitoring using machine-to-machine (M2M) technology, especially for patients treated at home or remote clinics. The end-to-end M2M will be built and a prediction system will be embedded as the novel feature.

REFERENCES

[1] 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).

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

[3] 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).

[4] Kementerian Kesehatan Republik Indonesia. Profil Kesehatan Indonesia 2012. Accessed via http://www. depkes. go. id (2013).

[5] Chen, M., Wan, J., González, S., Liao, X., & Leung, V. A survey of recent developments in home M2M networks. Communications Surveys & Tutorials, IEEE, 16(1), 98-114 (2014).

[6] Jung, S., Ahn, J. Y., Hwang, D. J., & Kim, S. An optimization scheme for M2M-based patient monitoring in ubiquitous healthcare domain. International Journal of Distributed Sensor Networks (2012).

[7] Turcu, C. E., & Turcu, C. O. Internet of things as key enabler for sustainable healthcare delivery. Procedia-Social and Behavioral Sciences, 73, 251-256 (2013).

[8] Fan, Z. M2M communications for E-health: Standards, enabling technologies, and research challenges. In 2012 6th International Symposium on Medical Information and Communication Technology (ISMICT) (pp. 1-4) (2012).

[9] 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).

[10] Han, J., Kamber, M., & Pei, J. Data mining: concepts and techniques. Elsevier (2011).

[11] Thuraisingham, B. A primer for understanding and applying data mining. IT Professional, 2(1), 28-31 (2000).

[12] Chaurasia, V., & Pal, S. Early prediction of heart diseases using data mining techniques. Carib. j. SciTech, 1, 208-217 (2013).

[13] Jarad, A., Katkar, R., Shaikh, A. R., & Salve, A. Intelligent Heart Disease Prediction System With MongoDB. International Journal of Emerging Trends & Technology in Computer Science, Volume 4, Issue 1, January-February 2015 (2015).

[14] Karthiga, G., Preethi, C., & Devi, R. Heart Disease Analysis System Using Data Mining Techniques. In 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14). IEEE (2014).

[15] 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).

[16] M. Akhbari, Y. Zare Mehrjerdi, H. Khademi Zare, A. Makui. A Novel Continuous KNN Prediction Algorithm to Improve Manufacturing Policies in a VMI Supply Chain. International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 27, No. 11, (November 2014) 1681-1690 (2014).

[17] M. R. Shafiee-Chafi, H. Gholizade-Narm. A Novel Fuzzy Based Method for Heart Rate Variability Prediction. International Journal of Engineering (IJE) Transactions A: Basics, Vol. 27, No. 7, 1041-1050 (2014).

[18] Hassanpour, H., and M. Mesbah. Newborn EEG seizure detection based on interspike space distribution in the time-frequency domain. International Journal of Engineering (IJE) Transactions A: Basics. 20.2: 137 (2007).

[19] Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., & Wets, G. (2000, August). A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. In Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 300-304). ACM.

[20] 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.

[21] 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.

[22] 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.

[23] Lin, R. H. (2009). An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine, 47(1), 53-62.

[24] Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM New York.

[25] D'Agostini, G. (1995). A multidimensional unfolding method based on Bayes' theorem. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 362(2), 487-498.

[26] Lavanya, D., & Rani, K. U. (2012). Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services (IJITCS), 2(1), 17-24.

[27] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[28] Khan, M., Ding, Q., & Perrizo, W. (2002). K-nearest neighbor classification on spatial data streams using p- trees. In Advances in Knowledge Discovery and Data Mining (pp. 517-528). Springer Berlin Heidelberg.

[29] Janosi, A., M.D. Heart Disease Data Set: Hungarian Institute of Cardiology. Accessed via: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (1988).

[30] Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. Data mining in bioinformatics using Weka. Bioinformatics 20(15), 2479-2481 (2004).

[31] Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In Convergence Information Technology, 2007 International Conference on (pp. 1541-1546). IEEE (2007).

[32] Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. arXiv Preprint arXiv:1409.0919.