Comparative Analysis of Classification Method for Wart Treatment Method

Genda Ananta Rahmat¹, Rifkie Primartha¹, Sukemi¹, Adi Wijaya²

¹Faculty of Computer Science, Sriwijaya University, Indonesia
²Informatics Engineering Department, Universitas MH Thamrin, Jakarta, Indonesia

E-mail: gendaanantarrahmat@gmail.com, rifkie@ilkom.unsri.ac.id, adiwjj@gmail.com, sukemi@ilkom.unsri.ac.id

Abstract. Wart disease is a benign tumor that grows in all parts of the body with varying amounts. There have been many methods of treatment of wart diseases that are found, but not much research leads to proving which treatments are best at treating wart disease. As a result, Patients sometimes need extra costs and time to cure wart disease because they have to be treated with more than one method. The purpose of this study was to compare the method of data mining classification to predict the best treatment of warts for patients between cryotherapy treatment which is the treatment of warts in general and the treatment of immunotherapy which is the latest treatment. Decision tree, random forest and k-nearest neighbor will be used as a classifier in predicting wart treatment in this study. From the research, the results obtained from the comparison of these three algorithms are amazing and promising with the best prediction accuracy on cryotherapy treatment achieved by the k-nearest neighbor algorithm of 95.66% while the best accuracy for immunotherapy treatment is achieved by random forest algorithm with an accuracy of 88.89%. We also predicted by combining the two datasets into 1 part, the result of the comparison for the three algorithms with the combined dataset it was found that the k-nearest neighbor algorithm was the best classifier to predict the wart treatment method with an accuracy of 88.03%.

1. Introduction

The Data Mining Algorithm (Knowledge Database Discovery) is used to find knowledge or patterns from a set of data. Data Mining algorithm has been widely used to analyze data in the field of medical science [1-10], information retrieval, market basket analysis and others in order to generate knowledge that can support decision making.

Classification, one of the five methods in data mining, is a commonly used method for predicting disease or the best treatment in health [1]. Decision tree [2] [3], random forest [4], naïve bayes [5], k-nearest neighbor, gradient boost tree are examples of algorithms developed in the classification method. With the many algorithms available today, various methods can be applied in one case to achieve the best results in extracting knowledge in areas such as medical science, for example in the prediction of chronic kidney disease [6].

In medical research, the combination of various classification algorithms and data reduction with the same dataset has also been widely performed [7]. For example, several studies conducted with pima Indians diabetes dataset [8] include the use of the C.45 algorithm with an accuracy of 81.27% [9], hybrid prediction model using simple k-means clustering and C4.5 algorithm with an accuracy of 92.38%, to use new proposed support vector machine with 100% classification accuracy [11].
On the other hand, research on the treatment of wart disease is a new field in medical science [12] which is still very minimal in research activities. Although there has been provided a public dataset to support research since 2018 [13] [14], Throughout our best knowledge in making this paper, there is only 1 study that discusses expert systems in predicting the treatment of wart disease [15] by creating expert systems with fuzzy rule algorithms -based system combined with ANFIS to predict whether a treatment is the best treatment to cure wart disease for patients. In this study, the accuracy achieved was 80.00% for cryotherapy treatment prediction and 83.33% for prediction with immunotherapy treatment. But in the field of data mining, there has been no recent research done even though the dataset has been disseminated by researchers since January 2018 ago.

In this study, we conducted a study comparing the analysis of the classification method, the Decision Tree, the Random Forest, and the k-nearest neighbor using immunotherapy datasets [13] and cryotherapy [14] as well as the combined datasets of the two previous datasets in predicting appropriate treatment for patients using the treatment of warts in general, cryotherapy, and the latest treatment of immunotherapy. The purpose of this research is to help researchers and other developers develop a better system in predicting the treatment of wart disease in order to save time and cost for patients with wart disease.

2. Material and Proposed Method
The dataset used in this study was obtained from donations in UCI public repository given by khozeimeh, fahime et al in January 2018. There are 2 donated datasets of cryotherapy [13] and immunotherapy [14]. The Cryotherapy dataset has 7 attributes (6 regular attributes, 1 label) with the number of instances reaching 90, while the Immunotherapy dataset has 8 attributes (7 regular attributes, 1 label) with the number of instances reaching 90.

The researcher will also create a new dataset by combining the two previous datasets. In this study, the two datasets were combined in order to test the robustness of the proposed method. Merging these two datasets is done only in cases that prove successful in each method. Then, the features that are not the same between datasets are not included in this merge. So that the new dataset resulting from this merger has dimensions of 118 cases with 71 case of immunotherapy and 47 case of cryotherapy. Table 1 below shows the feature details of the combined datasets.

| Feature Name                              | Value          |
|-------------------------------------------|----------------|
| Response of Treatment(Label)               | 1-Cryotherapy  |
|                                           | 2-Immunotherapy|
| Gender                                    | 1-Man          |
|                                           | 2-Woman        |
| Age(year)                                 | 15-67          |
| Time Elapsed before Treatment(month)       | 0-12           |
| The number of warts                       | 1-19           |
| Type of Warts                             | 1-Common       |
|                                           | 2-Plantar      |
|                                           | 3-Both         |
| Surface area of warts                     | 4-900          |

In this study, we propose research by applying various kinds of classification algorithms to the obtained dataset ([13] [14] and combined) in order to measure how well the classification algorithm predicts the correct treatment for patients with wart disease. As shown in Figure 1, Before the data is processed, Pre-processing data is performed using a forward selection algorithm to select the best feature to reduce the attribute. Furthermore, the data is divided into 10 parts using 10-fold validation where 9 parts of data will be used as training data and 1 part of the data will be used as testing up to 10 times
testing. The next stage is Information Selection ENN will be used to select the instance of the training dataset for optimal data training. After data passing the pre-processing step and data separation have been done by 10-fold validation, Application of training data of each classification algorithm (Decision Tree, Random Forest, and K-nearest neighbor) is done to make knowledge model. Once the model is obtained, test data is tested against the model that has been formed. This process will run as much as 3 times where each process will use different algorithms i.e. decision tree, random forest, and k-nearest neighbor. Figure 1 shows the detailed research flow in the flowchart.

In this research, process method will be evaluated using classifier effectiveness. Later on, each confusion matrix generated will be used to calculate the performance of each algorithm. Based on the confusion matrix, then the following calculation of the measurement will be done:

(i) Accuracy : classifier ability to select all instances that should be selected (true / TP) and reject all instances that should be rejected (wrong / TN).

\[
\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}
\]

(ii) Sensitivity : classifier ability in selecting all instances that should be selected (true / TP).

\[
\text{Sensitivity} = \frac{TP}{(TP + FN)}
\]

(iii) Specificity : classifier ability to reject all instances that should be rejected (wrong / TN).

\[
\text{Specificity} = \frac{TN}{(TN + FP)}
\]

(iv) Informedness : value trade off between sensitivity and specificity.

\[
\text{Informedness} = (\text{Sensitivity} + \text{Specificity}) - 1
\]

As well as ROC Curve measurements will also be used in this study.

![Flowchart of Proposed Method](image)

**Figure 1.** Flowchart of Proposed Method

In the end, after obtaining the results of each algorithm, anova test will be performed to see if there is a significant difference between the results of each algorithm. Anova results will be considered significant if probability < \(\alpha\), with \(\alpha\) of 0.5.
3. Experimental Result

The study was conducted using computer calculations with the details of the Intel (R) Core (TM) i7-6500 CPU @ 2.5Ghz (4 CPUs), - 2.6Ghz, DDR3 8 GB RAM, Windows 10 Home 64 bit as operating system and Rapid Miner Version 8.1.0.0 as a data analysis tool. Rapid miner will also produce final calculations such as accuracy and confusion matrix.

The test was performed in 3 stages, test 1 was performed with cryotherapy data [13], test 2 was performed with immunotherapy data [14], and test 3 was performed with synthesized data from both datasets [13] [14]. Measurement The classification performance of each test for each decision Tree, Naïve Bayes, and Random Forest algorithm that has been executed from all datasets is represented in Table 2, Table 3 and Table 7. Table 5 shows the results of the selected features for each test.

| Algorithm                | Accuracy | Sensitivity | Specificity | AUC Value | Informdness |
|--------------------------|----------|-------------|-------------|-----------|-------------|
| Decision Tree            | 92.22%   | 0.952       | 0.895       | 0.628     | 0.847       |
| Random Forest            | 94.44%   | **0.976**   | 0.916       | **0.960** | 0.892       |
| K-Nearest Neighbor       | **95.56%** | 0.928       | **0.979**   | 0.500     | **0.907**   |

| Algorithm                | Accuracy | Sensitivity | Specificity | AUC Value | Informdness |
|--------------------------|----------|-------------|-------------|-----------|-------------|
| Decision Tree            | 87.78%   | **0.971**   | 0.526       | 0.714     | 0.497       |
| Random Forest            | **88.89%** | **0.971**   | 0.578       | 0.657     | 0.549       |
| K-Nearest Neighbor       | 87.78%   | 0.929       | **0.684**   | **0.729** | **0.613**   |

| Expert System            | Accuracy | sensitivity | specificity | AUC Value |
|--------------------------|----------|-------------|-------------|-----------|
| K-Nearest Neighbor       | **95.56%** | 0.928       | **0.979**   | 0.500     |
| Random Forest            | 94.44%   | **0.976**   | 0.916       | **0.960** |
| Decision Tree            | 92.22%   | 0.952       | 0.895       | 0.628     |
| Expert System Using Fuzzy Rule-Based System[15] | 80.00% | 0.820 | 0.770 | 0.902 |

| Expert System            | Accuracy | sensitivity | specificity | AUC Value |
|--------------------------|----------|-------------|-------------|-----------|
| Random Forest            | **88.89%** | **0.971**   | 0.578       | 0.657     |
| K-Nearest Neighbor       | 87.78%   | 0.929       | 0.681       | 0.729     |
| Decision Tree            | 87.77%   | **0.971**   | 0.526       | 0.714     |
| Expert System Using Fuzzy Rule-Based System[15] | 80.00% | 0.820 | 0.770 | **0.902** |

In Tables 1 and 2, each classifier provides a very satisfied performance in predicting the right treatment for patients with wart disease. In the test result 1 in Table 2, k-nearest neighbor is the best algorithm to achieve the accuracy(sensitivity; specificity) of 95.56%(0.928; 0979) with value of k=1, value of k for enn = 8 followed by the random forest that achieved 94.44%(0.976; 0.916) accuracy with criterion used is information gain, number of tree = 9, value of k for enn = 4 and the last decision tree reached accuracy 92.22%(0.952; 0.895) with criterion used is information, value of k for enn = 4. When compared to the performance results of expert systems [15] in table 4 using the cryotherapy dataset, the
three algorithms used in test 1 far exceed the results of previous studies in terms of accuracy, specificity, and auc where k-nearest neighbor algorithm has the values best of accuracy, while for the highest sensitivity is achieved by random forest and the highest specificity is achieved by k-nearest neighbor.

While the result of test 2 in Table 3 shows that random forest algorithm has achieved the best accuracy (sensitivity; specificity) of 88.89% (0.971; 578) with criterion used is gini index, number of tree = 10, value of k for enn = 3 followed by decision tree and k-nearest neighbor which achieves same 87.88% accuracy with details for knn value of k=9 knn, value of k for enn = 5 and for decision tree criterion used is gini index and value of k for enn = 9. When compared to the performance results of expert systems [15] in Table 5 using immunotherapy datasets, the three algorithms used in this study far exceed the results of previous studies in terms of accuracy and sensitivity of classification random forest is best, for the highest value auc is achieved by the k-nearest neighbor whereas expert systems [15] in earlier studies were only capable of beating only in terms of the value of specificity and auc.

| Test | Dataset   | Decision Tree          | Random Forest          | K-nearest Neighbor          |
|------|-----------|------------------------|------------------------|-----------------------------|
| 1    | Cryotherapy| - Area, - Type, - Age, - Result of Treatment | - Area, - Age, - Result of Treatment | - Area, - Type, - Age, - Time, - Gender, - Result of Treatment |
| 2    | Immunotherapy| - Time, - Age, - Induration, - Number, - Gender, - Result of Treatment | - Time, - Age, - Result of Treatment | - Time, - Result of Treatment |
| 3    | Combined  | - Area, - Type, - Age, - Result of Treatment | - Area, - Age, - Result of Treatment | - Area, - Type, - Age, - Time, - Sex, - Result of Treatment |

Table 6 shows the results of the data reduction algorithm using forward selection. In the first test, it appears that the attributes on the dataset are reduced in each of the algorithms tested. In the decision tree algorithm, the attributes of gender, time, and number are omitted while in the random forest algorithm, the attribute type, gender, time, and number are also omitted, and in the knn algorithm, only the attribute number is omitted. The same results are also shown in the second test where the decision tree algorithm, type and area attributes are omitted while in the random forest induration algorithm, gender, type, number and area are omitted and in the knn algorithm, attributes are omitted i.e. age, induration, number, gender, type and area. For the third test, on the decision tree algorithm, the attributes that are set aside are gender, time and number while in the random forest algorithm, attributes that are omitted are gender, time, number and type and in the knn algorithm, the attribute is omitted only number. The result of attribute reduction for each test in each algorithm is shown in the table below.
Table VII. Comparison Performance Each Algorithm with Combined Dataset

| Algorithm                | Accuracy | Sensitivity | Specificity | AUC curve | Informness |
|--------------------------|----------|-------------|-------------|-----------|------------|
| Decision Tree            | 73.79%   | 0.638       | 0.802       | 0.750     | 0.440      |
| Random Forest            | 84.17%   | 0.765       | 0.887       | 0.851     | 0.652      |
| K-Nearest Neighbor       | 88.03%   | 0.936       | 0.845       | 0.500     | 0.781      |

From the test results using the synthesis dataset by combining the two datasets in order to predict more practically without being required to predict patient data one by one to each dataset, K-nearest neighbor is the best algorithm in predicting the right treatment method for patients with wart disease. The k-nearest neighbor algorithm achieved 88.03% accuracy, sensitivity of 0.936, specificity of 0.845, and a roc curve value of 0.500 with value of k=1 and value of k for enn = 6.

In the second rank, the random forest algorithm can predict with accuracy reaching 84.17%, sensitivity of 0.765, specificity of 0.887, and the value of the curve of roc is 0.851 with criterion used is information gain, number of tree = 17, value of k for enn = 4.

While decision tree algorithm only managed to predict the right treatment for patients with wart disease with an accuracy of 73.79%, sensitivity of 0.638, specificity of 0.802, the value of the curve roc of 0.750 with criterion used is gain ration, value of k for enn = 7.

Table VIII. Anova Test Result

| ANOVA | f   | prob | Alpha |
|-------|-----|------|-------|
| Test 1| 0.467 | 0.632 | 0.05  |
| Test 2| 0.056 | 0.954 | 0.05  |
| Test 3| 3.885 | 0.033 | 0.05  |

From the results of anova test in Table 8, In testing the difference of accuracy result from each algorithm got that difference at test 1 and 2 is not significant. While in test 3 using the combined dataset, it appears that probability < alpha so that it can be deduced that there is a significant difference of each result accuracy achieved.

In the end, the research we have done can provide a higher level of accuracy in predicting the best treatment method for patients with wart disease by applying the method of data mining classification compared to previous studies [15] by creating an expert system based on the fuzzy rule-based system combined with ANFIS.

4. Conclusion and Future Work
In this study, some classification algorithm has been applied in 3 experimental simulations to measure the performance of algorithms in predicting the best treatment method for patients with wart disease. After further experiments, it has been proved that the 3 algorithms used in this study can produce better accuracy from expert systems that have been made in previous studies [15]. In test 3 using the synthesis dataset, the accuracy of the k-nearest neighbor algorithm is the best algorithm in comparison with an accuracy of 88.01%. From ANOVA statistic test results also obtained the probability value < alpha in test 3, which proves that there is a significant difference from the performance of each algorithm obtained.

Future work can be focused on reducing features to improve computing efficiency if future datasets increase without reducing the accuracy of the classification algorithm that has been used.

References
[1] L. Kodali, A.S. Adusumilli, P. Doreti, D. T.Renuga, A. Umamakeswari, "Performance analysis of various data mining techniques in the prediction of heart disease", Ind. Jou. Sci. Tech, vol 8, 2015.
[2] T. Maryam, E. Habibollah, S.K. Maryam, A.Z. Alireza, E. Mahmoud, S. Mohammad, N. Mohsen,
M.R.P. Seyed, A.F. Gordon, G.M. Mobarak, "The application of a decision tree to establish the parameters associated with hypertension," Comp. Meth. Prog. Biomed, vol 139, February 2017.

[3] P. Sajida, S. Muhammad, G. Aziz, K. Karim, "Performance analysis of data mining classification to predict diabetes," Proc. Comp. Sci, vol 82, 2016.

[4] M. Zerina, S. Abdulhamit, "Cognitive heart failure detection using random forest classifier," Comp. Meth. Prog. Biomed, vol 130, 2016.

[5] H. Mohammad R, H. Naeemul, K. Rayan, K. Young-Tae, S.M. Iqbal, "Classification of cancer cells using computational analysis of dynamic morphology," Comp. Meth. Prog. Biomed, vol 156, 2018.

[6] R. L. Jerlin, P. Eswaran, "Generating comparative analysis of early stage prediction of chronic kidney disease," Int. Jou. Mod. Eng. Res(IJMER), vol 5, July 2015.

[7] J. Divya, S. Vijendra, "Feature selection and classification systems for chronic disease prediction: a review," Egy. Inf. Jou, 2018.

[8] Kayaer K, Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks" Pro. Int. Conf. Artif. Neu. Net Neu. Info. Proces(ICANN/ICONIP), 2003.

[9] Purushottam, S. Kanak, S. Richa, "Diabetes mellitus prediction system evaluation using c4.5 rules and partial tree" Int. Conf. Relia. Info. Tech. Opti(ICITRO), 2015.

[10] B.M Patil, R.C Joshi, T. Durga, "Hybrid prediction model for type-2 diabetic patients" Exp. Sys. App, vol 37, 2010.

[11] G. Emre, K. Erdal, "A new adaptive support vector machine for diagnosis of diseases" Exp. Sys, vol 31, 2013.

[12] K. Fahime, J.A Farahzad, M.O Yaghoub, J. Majid, T. Shahrzad, A. Roohallah, L. Pouran, "Intralesional immunotherapy compared to cyrotherapy in treatment of warts," Int. Jou. Derma, vol 56, 2017.

[13] K. Fahime, Cryotherapy Dataset, “UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].” Mashad, Iran: Faculty of Medicine, Mashad University of Medical Sciences, 2018.

[14] K. Fahime, Immunotherapy Dataset, “UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].” Mashad, Iran: Faculty of Medicine, Mashad University of Medical Sciences, 2018.

[15] K. Fahime, A. Roohallah, R. Mohamad, K. Abbas, L. Pouran, N. Saeid, "An expert system for selecting wart treatment method," Comp. Bio. Medic, vol 81, 2017.