Performance Comparison of Machine Learning Algorithms for Classification of Chronic Kidney Disease (CKD)

Azian Azamimi Abdullah¹, Syazwani Adli Hafidz¹ and Wan Khairunizam¹

¹ School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia

Abstract. Kidney is one of the vital organs in a human body while ironically, chronic kidney disease (CKD) is one of the main causes of death in the world. Due to the low rate of loss of kidney function, the disease is often overlooked until it is in a really bad condition. Dysfunctional kidney may lead to accumulation of wastes in blood which would affect several other systems and functions of the body such as blood pressure, red blood cell production, vitamin D and bone health. Machine learning algorithms can help in classifying the patients who have CKD or not. Even though several studies have been made to classify CKD on patients using machine-learning tool, not many researchers perform pre-processing and feature selection technique to obtain quality and dependable result. Machine learning used with feature selection techniques are shown to have better and more dependable result. In this study, feature selection methods such as Random Forest feature selection, forward selection, forward exhaustive selection, backward selection and backward exhaustive selection were identified and evaluated. Then, machine learning classifiers such as Random Forest, Linear and Radial SVM, Naïve Bayes and Logistic Regression were implemented. Lastly, the performance of each machine-learning model was evaluated in terms of accuracy, sensitivity, specificity and AUC score. The results showed that Random Forest classifier with Random Forest feature selection is the most suitable machine learning model for classification of CKD as it has the highest accuracy, sensitivity, specificity and AUC with 98.825%, 98.04%, 100% and 98.9% respectively which outperformed other classifiers.

1. Introduction

Kidney is one of the important organs in a human body. They filter blood in system, removing toxic and wastes through urine and maintain the balance of bodily fluid in one’s body. Chronic Kidney Disease (CKD) is when the Glomerular Filtration rate (GFR) decreases 60ml/min/1.73m² in the duration of three months or longer. Due to low rate of loss of kidney function, the disease is often overlooked until it is in a really bad condition [1]. In 1990, a study showed that death caused by CKD had increased from 409,000 deaths in 1990 to 956,000 deaths in 2013 [2]. As of 2010, the commonness of people receiving renal replacement treatment such as dialysis is shown to be 96.8% in 28 Asian countries [3].

Feature selection is one of the important steps priors to develop machine-learning algorithm as it may increase the accuracy of the algorithm. Potharaju et al [4] used SVM classifier to identify significant features and resulted with high average accuracy of 93.7%. Polat et al [5] used both...
Wrapper and Filter method to be used with SVM classifier which resulted in 98.5% highest obtained accuracy and 97.75% of the lowest obtained accuracy. Zhou et al [6], Misir et al [7], Chetty et al [8] and Tazin et al [9] all had proved that feature selection increases the accuracy of a machine learning model. This study will obtain the significant features contributing to the performance of machine learning before implementing machine learning models.

2. Methodology

This section presents the general workflow of this study, including the data extraction, data preprocessing, data visualization, feature selection and classification.

2.1. Data Extraction

The dataset is extracted from UCI database, which consists of 400 subjects, 24 features and one output collected from Apollo Hospitals in Tamilnadu [10]. The last column is the output whether the subject has CKD or not.

2.2. Data Pre-processing

There are some missing values, redundant information and impossible data. If the data is included in the algorithm, the result may come out as false and affects the accuracy of the result. Therefore, with data pre-processing technique the raw data extracted will be processed into clean data which will give better and more accurate output. Data pre-processing consists of four stages which are data cleaning, data integration, data transformation and data reduction.

2.3. Data Exploration/Visualization

The data given is divided into two types, which are numerical data and nominal data. Therefore, the visualisation of data uses two types, which are histogram for numerical data, and countplot, which is the categorical count chart.

2.4. Feature Selection

In this study, five feature selection methods will be used. One from embedded approach, which is Random Forest feature selection, and four from the Wrapper approach, which are forward selection, forward exhaustive selection, backward selection and backward exhaustive selection.

2.4.1. Forward Selection

Forward selection method is a repetitive approach which begins by using no features in the algorithm. Then, the feature will be added one by one in each repetition in order to discover the feature which improves the algorithm the most. The process repeats until the inclusion of the feature does not give any effect to the performance of model.

2.4.2. Backward Selection

Backward selection method begins with the initial features of the dataset. In each repetition, the feature will be dropped one by one starting with the least significant feature that does not improve the algorithm. The repetition will stop when the remaining features have the same significance.

2.4.3. Exhaustive Selection

Exhaustive feature selection evaluates the performance of all feature combinations from the best subsets selected by forward and backward selection methods. It will have a much smaller subset of features compared to both forward and backward selection methods.

2.4.4. Embedded Approach

The embedded method performs both feature selection and machine learning algorithm simultaneously. This method is less extensive in terms of computation as compared to Wrapper method. The embedded method that will be used in this project is using the Random Forest algorithm.

2.5. Machine Learning Model
In this project, there will be two rounds of prediction to be done which are prediction using all 24 features of the dataset and prediction using the selected features after using the feature selection technique. Supervised machine learning is used as there is one targeted output; whether a person will have CKD or not. The examples of supervised machine learning that will be used is Random Forest, Support Vector Machine (SVM), Naïve Bayes and Logistic Regression.

2.5.1. Random Forest Classifier. Random Forest is one of the most used machine learning for classification. Basically, it accumulates numbers of decision trees and combines them together for a more accurate and stable prediction. Random Forest adds additional unpredictability to the algorithm, while the branches continue to split. The classifier identifies the most significant feature among random subset rather than identifying them while splitting nodes. This gives various kinds of result which makes it a better algorithm. Thus, Random Forest considers random features to split a node. Also, it can handle big data smoothly and processes efficiently without any data deletion [11].

2.5.2. Support Vector Machine. Support Vector Machine (SVM) is a supervised machine learning algorithm solves both classification and regression tasks. An SVM model categorizes new examples into another group or the other which makes it a non-probabilistic binary linear classifier [12]. This model plots each data as a point in n-dimensional space where the value of each feature is the coordinate on the plane. Next, the two divisions are divided by producing the hyperplane which divides the two of them. This means that this method of machine learning produces an optimum hyperplane from the coordinated data which classifies new examples.

2.5.3. Naïve Bayes. The Naïve Bayes method is based on Bayes’ theorem with the independence assumptions between predictors. A Naïve Bayes algorithm can be made easily, where only simple parameter estimation needed. This is very advantageous when used for big datasets. Even if it is simple, this machine learning method excels and is popularly used because it always exceeds the more difficult methods. If the Naïve Bayes class conditional independence actually performs, a Naïve Bayes classifier will assemble quicker than segregating algorithms like logistic regression [13].

2.5.4. Logistic Regression. Logistic regression analyses prediction where it is used to explain data and describe the correlation of a dependent binary variable with another variable which could be nominal, ordinal or interval. This machine learning model aims to determine the best fitting model to explain the correlation between the dichotomous feature of interest and a set of independent variables. Logistic regression is suitable for this project as the dependent variable, which is also the outcome, is a binary variable (CKD/NON-CKD) and there is a set of 24 features.

2.6. Evaluation of Machine Learning Model
The algorithms will be evaluated in terms of accuracy, sensitivity and specificity. To evaluate, confusion matrix, receiver operating characteristics (ROC) and area under curve (AUC) will be used.

2.6.1. Confusion Matrix. A confusion matrix is a table which is frequently used to evaluate the performance of an algorithm on a group of test data where the true values are known. This allows more comprehensive analysis than just accuracy. Accuracy is not a dependable parameter of the actual practice of an algorithm as it can result to incorrect output if the dataset is unbalanced.

2.6.2. Receiver Operating Characteristics (ROC) and Area Under Curve (AUC). An ROC curve is a graph, which illustrate the performance of an algorithm at all classification limits. The curve plots two parameters which are sensitivity versus 1-specificity at various classification limits. Reducing the classification limit categorizes more data as positive, thus rising both True Positive and False positives. AUC is the area under the ROC curve. AUC gives the cumulative value of performance in all viable classification limits. It explains the ability of a machine learning model to distinguish between classes. AUC also can be explained as the possibility where the algorithm places a random positive example at a higher rank than a random negative example.
3. Results and Discussion

Data pre-processing needs to be done to the raw data as it there are many missing values. The missing values are labelled as NaN and it is replaced with 0. Also, the nominal data will be replaced with 1 and 0 according to its good or bad. 1 is good for health and 0 is bad for health.

The Random Forest classifier is one of the embedded methods used to choose the features alongside 10-fold cross validations. The 10-fold cross validation is used to train and test part of the data with the accuracy to determine the most important features. For Wrapper approach, forward selection, backward selection and exhaustive selection is used. The dataset is separated into two parts, one for training and another for testing. 320 samples and 24 features are used for training while 80 samples and 24 features are used for testing. Table 1 is the features selected by all five features selection methods.

Table 1. Features Selected by Random Forest Features Selection, Forward Selection, Forward Exhaustive Selection, Backward Selection and Backward Exhaustive Selection.

| Random Forest Feature Selection | Forward Selection | Forward Exhaustive Selection | Backward Selection | Backward Exhaustive Selection |
|---------------------------------|-------------------|-----------------------------|-------------------|-----------------------------|
| Specific Gravity                | Specific Gravity  | Specific Gravity            | Red Blood Cell    | Serum Creatinine            |
| Albumin                         | Albumin           | Albumin                     | Anaemia           | Appetite                    |
| Sugar                           | Sugar             | Haemoglobin                 | Serum Creatinine  | Diabetes Mellitus           |
| Blood Glucose Random            | Red Blood Cell    | Appetite                    | White Blood Cell  | Packed Cell Volume          |
| Serum Creatinine                | Bacteria          |                             | Red Blood Cell    | Red Blood Cell Count        |
| Potassium                       | Potassium         |                             |                   | Appetite                    |
| Packed Cell Volume              | Haemoglobin       |                             | Diabetes Mellitus |                             |
| White Blood Cell Count          | Packed Cell Volume|                             | Packed Cell Volume|                             |
| Red Blood Cell Count            | Appetite          |                             |                   | Pedal Edema                 |
| Diabetes Mellitus               | Pedal Edema       |                             | Blood Glucose Random|                             |

Machine learning classifiers are then implemented towards the selected features and also towards all features (without any feature selection). The machine learning models are then evaluated in terms of accuracy, sensitivity and specificity. Table 2 shows the highest result from each classifier.
Table 2. Highest Accuracy, Sensitivity and Specificity from Each Classifier

| Model                                                      | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|-----------------------------------------------------------|--------------|-----------------|-----------------|
| Random Forest Classifier with Random Forest Feature Selection | 98.825       | 98.040          | 100.000         |
| Linear SVM with All Features                              | 97.500       | 98.000          | 96.670          |
| Radial SVM with Forward Selection                         | 95.000       | 94.000          | 96.670          |
| Radial SVM with Forward Exhaustive Selection              | 95.000       | 94.000          | 96.670          |
| Naïve Bayes with Backward Selection                       | 97.500       | 100.000         | 90.000          |
| Logistic Regression with Random Forest Feature Selection   | 96.250       | 94.000          | 100.000         |

From Table 2, Random Forest classifier with Random Forest feature selection has the highest accuracy of 98.825% followed by Linear SVM with all features, Naïve Bayes with backward selection, Logistic Regression with Random Forest feature selection and Radial SVM with both forward selection and forward exhaustive selection with 97.5%, 97.5%, 96.25%, 95% and 95% respectively. Naïve Bayes with backward selection has the highest sensitivity of 100% followed by Random Forest classifier with Random forest feature selection, Linear SVM with all features, Radial SVM with forward selection and forward exhaustive selection and Logistic Regression with Random Forest feature with 98.04%, 98%, 94%, 94% and 94% respectively. Random Forest classifier with Random Forest feature selection and Naïve Bayes with backward selection have the highest specificity with 100% followed by Linear SVM with all features, Radial SVM with forward selection and forward exhaustive selection with 96.67% and Naïve Bayes with backward selection with 90%.

When comparing the features involved in all the models above, it is found that there are a few features that repetitively used in various models. The most common feature selection used in Random Forest feature selection, forward selection and forward exhaustive selection are specific gravity and albumin. Appetite is a common feature between forward selection, forward feature selection and backward selection. Blood glucose random, serum creatinine, packed cell volume, white blood cell count, red blood cell count and diabetes mellitus are the common features used in Random Forest feature selection and backward selection.

The values of AUC were also compared to evaluate the performance of machine learning models. Table 4 shows the highest AUC values from each classifier

Table 3. Highest AUC Score for Each Classifier

| Model                                                      | AUC (%) |
|-----------------------------------------------------------|---------|
| Random Forest Classifier with All Features                | 98.90   |
| Random Forest Classifier with Forward Exhaustive Selection| 98.90   |
| Linear SVM with All Features                              | 97.33   |
| Radial SVM with Forward Selection                         | 95.33   |
| Radial SVM with Forward Exhaustive Selection              | 95.33   |
| Naïve Bayes with All Features                             | 97.33   |
Random Forest classifier with all features and Random Forest with Forward Exhaustive Selection have the highest AUC value which is 98.9%. Both Linear SVM with all features and Naïve Bayes with all features come second with 97.33% followed by both Logistic Regression with all features and Logistic Regression with Random Forest feature selection. Finally, Radial SVM with forward selection and Radial SVM with forward exhaustive selection come last with 95.33%.

For CKD classification, Random Forest classifier with Random Forest feature selection is the best in terms of accuracy (98.825%) and specificity (100%). Also, Random Forest classifier with all features and Random Forest with Forward Exhaustive Selection is the best in terms of AUC (98.9%) values and specificity (100%).

4. Conclusion
This study has identified and evaluated five feature selections; Random Forest feature selection, forward selection, forward exhaustive selection, backward selection and backward exhaustive selection. Also, four machine learning algorithms have been selected to classify CKD which are Random Forest classifier, Linear and Radial SVM, Naïve Bayes and Logistic Regression. The performance of the machine learning algorithms has also evaluated in terms of accuracy, sensitivity, specificity and AUC. Overall, Random Forest classifier with Random Forest feature selection performed the best in terms of accuracy (98.825%), sensitivity (98.040%) and specificity (100%). For future work, the application of classification of CKD can be further developed with the aid of graphical user interface (GUI) and can be used as computer-aided system for the diagnosis of CKD. Also, collaborations can be done with Malaysian hospitals to obtain CKD data from patients in order to improve the healthcare system in Malaysia.

References
[1] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson 2017 The Lancet
[2] GBD 2013 Mortality and Causes of Death Collaborators 2015 Lancet (London, England)
[3] T. Liyanage, T. Ninomiya, V. Jha, B. Neal, H. M. Patrice, I. Okpechi, M. Zhao, J. Lv, A. X. Garg, J. Knight, A. Rodgers, M. Gallagher, S. Kotwal, A. Cass, and V. Perkovic 2015 Lancet 385 1975–1982
[4] S. Prasad Potharaju and M. Geetha 2018 Artic. Int. J. Eng. Technol. 7 p 191–194
[5] H. Polat, & Homay, D. Mehr, and A. Cetin 2017 J. Med. Syst. 41
[6] L.-T. Zhou, Y.-H. Cao, L.-L. Lv, K.-L. Ma, P.-S. Chen, H.-F. Ni, X.-D. Lei, and B.-C. Liu 2017 Sci. Rep. 7 39832
[7] R. Misir, M. Mitra, and R. K. Samanta 2017 J. Pathol. Inform. 8 p 24
[8] N. Chetty, K. S. Vaisla, and S. D. Sudarsan 2015 International Conference on Computing, Communication and Security (ICCCS) p 1–6
[9] N. Tazin, S. A. Sabab, and M. T. Chowdhury 2016 International Conference on Medical Engineering, Health Informatics and Technology (Meditec) p 1–6
[10] K. Bache and M. Lichman 2013 University of California Irvine School of Information 14/8 p 0
[11] G. Biau and E. Scornet 2016 Test 25 p 197–227
[12] P. Liu, K.-K. R. Choo, L. Wang, and F. Huang 2017 Soft Comput. 21 p 7053–7065
[13] S. S. Y. Ng, Y. Xing, and K. L. Tsui 2014 Appl. Energy 118 p 114–123