

# Chronic Kidney Disease Diagnosis System using Sequential Backward Feature Selection and Artificial Neural Network

Siti Noor Chotimah<sup>1\*</sup>, Budi Warsito<sup>2</sup>, and Bayu Surarso<sup>3</sup>

<sup>1</sup>Master Program of Information System, Postgraduate School, Diponegoro University, Semarang

<sup>2</sup>Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang

<sup>3</sup>Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang

**Abstract.** The number of factors that can be categorized into the diagnosis of Chronic Kidney Disease (CKD) at an early stage makes information about the diagnosis of the disease divided into information that has many influences and has little influence. This study aims to select diagnoses in medical records with the most influential information on chronic kidney disease. The first step is to select a diagnosis with much influence by implementing the Sequential Backward Feature Selection (SBFS). This algorithm eliminates features that are considered to have little influence when compared to other features. In the second step, the features of the best diagnoses are used as input to the Artificial Neural Network (ANN) classification algorithm. The results obtained from this study are information in the form of the best diagnoses that have much influence on chronic kidney disease and the accuracy results based on the selected diagnoses. Based on the study results, 15 features are considered the best of the 18 features used to achieve 88% accuracy results. Compared with conventional methods, this method still requires consideration from the medical staff because it is not a final diagnosis for patients.

## 1 Introduction

The increase in the number of patients suffering from Chronic Kidney Disease (CKD), which continues to occur every year, makes this disease one of the diseases included in a global public health problem. The level of the condition in patients with chronic kidney disease varies. Poor handling and lack of public awareness about this disease can lead to more severe kidney damage in the future, leading to death.

When the disease has been acknowledged and detected early, the risk of complications and the risk of worsening the condition of patients suffering from these diseases can be reduced. Medical treatment will immediately be carried out based on the early diagnosis that occurred in the patient. *Diagnosis* is an activity that has to do with collaboration between clinical reasoning and information gathering to determine health problems experienced by patients [1]. The many factors that can be categorized into the diagnosis of chronic kidney

---

\* Corresponding author: [imasitinoorc3424@gmail.com](mailto:imasitinoorc3424@gmail.com)

disease at an early stage make the information about the disease diagnosis divided into information that has a lot of influence and little influence.

There have been many previous studies regarding the diagnosis and/or related to chronic kidney disease, whether using text data or data in images. For example, one study of chronic kidney disease prediction compares the predictions of the Neural Network (NN) algorithm and Support Vector Machine (SVM) algorithm [2]. However, the main focus of this study is to select the diagnosis that most influences chronic kidney disease from the medical records that contain much information from patients. The method applied in this study is divided into two steps.

The first step is to select disease diagnoses with many influences by implementing Sequential Feature Selection (SFS). This method can evaluate the features that are not selected based on the most influential features [3], which relates to disease diagnosis. The type of SFS used in this study is the Sequential Backward Feature Selection (SBFS) type. The second step is to use the best diagnostic results to input the Artificial Neural Network (ANN) classification algorithm. The results obtained from this study are information of the best diagnoses that have much influence on chronic kidney disease and the accuracy results based on the selected diagnoses.

In this study, section 2 discusses the methods used in the study. Section 3 discusses the implementation of the study conducted through workflows and implementation of the methods used. Section 4 discusses the results obtained and discusses it. While in section 5 contains the conclusions of the study conducted.

## **2 Method**

### **2.1 Feature Selection**

Feature Selection is a technique in data pre-processing. It is often used in data mining applications, which helps reduce data by removing features that are considered unimportant from the dataset used [4]. The use of feature selection is considered to optimize the performance of the classification algorithm used in the study [5]. There are several advantages of feature selection in its application [5]: (1) Reducing noise in the dataset; (2) Helping the understanding of the final model in the classification algorithm; (3) Helping to update the model.

Traditional and modern approaches to feature selection to be applied in machine learning are divided into four categories as follows [4]: (1) Filter method, which filters out features before implementing any machine learning algorithm; (2) Wrapper method, which selects the features that are considered the best considering the machine learning algorithms used; (3) Embedded method, which measures the "usability" of a feature subset as well as makes feature selection; (4) Hybrid method, which combines one or more feature selection approaches to obtain the advantages of other methods.

### **2.2 Sequential Backward Feature Selection**

Sequential Feature Selection is a method in feature selection that can evaluate features that are not selected by referring to the selected features [3]. Types of strategies for using Sequential Feature Selection include [3]: (1) Sequential Forward Feature Selection (SFFS), which selects features based on "forward" sequential search strategy; (2) Sequential Backward Feature Selection (SBFS), which works the opposite of the SFFS method; (3) Plus-L Minus-R Selection (LRS), which is considered as an integration between the SFFS and SBFS methods.

Sequential Backward Feature Selection or also known as backward elimination, is a method that aims to reduce the number of features or dimensions of the data used [6]. Although this algorithm has a weakness which is not able to re-evaluate the usefulness of the feature after the feature is removed [3], the backward selection is considered the most suitable for use in the application of Artificial Neural Networks (ANN) [6]. The SBFS algorithm was chosen because this algorithm is included in the greedy wrapper method, which can choose a search path based on the direction that shows the best value of the features used in the study [7].

### 2.3 Artificial Neural Network

Artificial neural network (ANN) is a method that represents the most basic level to imitate the type of non-linear learning that occurs in the brain's neuron network. [8]. This method is based on how the brain works to establish an authentic connection [2]. The use of ANN has the advantage that the network formed can learn to overcome an uninformative example from existing data by requiring a longer training time than the decision tree [8].

The Backpropagation Algorithm is one of the general algorithms used in the application of the ANN method [9]. In use, there are three stages in the backpropagation algorithm: the feedforward stage, the backpropagation stage, and the weight adjustment stage [10]. First, the feedforward stage uses the following equation:

$$z\_in_j = v_{0j} + \sum_i x_i v_{ij} \quad (1)$$

where  $i$  and  $j$  are neurons in the input layer and the hidden layer,  $v_{0j}$  is the bias that exists in the input layer,  $x_i$  is the weights on  $i$ .  $v_{ij}$  is the value of the inputs on  $i$  to  $j$ .

Next, apply the activation function using the following equation:

$$z_j = f(z\_in_j) = \frac{1}{1 + \exp^{-z\_in_j}} \quad (2)$$

where  $f(z\_in_j)$  is the activation value contained in the hidden layer to the output and  $\exp$  is a "natural" number.

To calculate the other input, could use the following equation:

$$y\_in_k = w_{0k} + \sum_j z_j w_{jk} \quad (3)$$

where  $w_{0k}$  is the hidden layer bias,  $z_j$  is the value of the activation function,  $j$  is neuron,  $w_{jk}$  is the weight of  $j$  to  $k$  on output. Then perform the activation function again using the following equation:

$$y_k = f(y\_in_k) = \frac{1}{1 + \exp^{-y\_in_k}} \quad (4)$$

where  $f(y\_in_k)$  is the activation value contained in the hidden layer to the output. Second, the backpropagation stage using the following equation:

$$\delta_k = (t_k - y_k) f'(y\_in_k) \quad (5)$$

where  $\delta_k$  is the error value,  $y_k$  is the unit of output and  $t_k$  is the target pattern. Next, calculate the weight correction using the following equation:

$$\Delta w_{jk} = \alpha \delta_k z_j \quad (6)$$

where  $\Delta w_{jk}$  is the amount of change value to add to the old updated weight,  $\alpha$  is the learning rate. Then calculate the bias correction using the following equation:

$$\Delta w_{0k} = \alpha \delta_k \quad (7)$$

where  $\Delta w_{0k}$  is the amount of change value to add to the old updated weight. Next, calculate the input delta using the following equation:

$$\delta_{in_j} = \sum_{k=1}^m \delta_k w_{jk} \quad (8)$$

The results obtained from the calculation of the input delta multiplied by the derivative of the activation function, this process can be calculated using the following equation:

$$\delta_j = \delta_{in_j} f'(z_{in_j}) \quad (9)$$

where  $\delta_j$  is the error value.

Third, the weight adjustment stage using the following equation:

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk} \quad (10)$$

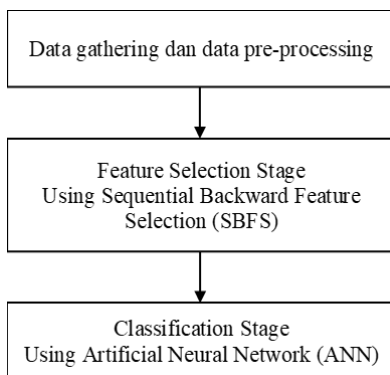
$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij} \quad (11)$$

where  $w_{jk}(new)$  is the new weight between the input layer and the hidden layer that will be searched for,  $v_{ij}(new)$  is the new weight between the hidden layer and the output layer to be searched for,  $w_{jk}(old)$  and  $v_{ij}(old)$  are updated old weights.

## 3 Implementation

### 3.1 Workflow

The steps in this study are divided into two steps. The first step is to select a diagnosis that has a lot of influence by implement the Sequential Backward Feature Selection (SBFS). This algorithm eliminates features that are considered to have little influence when compared to other features. While in the second step, the features consisting of the best diagnoses used as input to the Artificial Neural Network (ANN) classification algorithm. The explanation can be illustrated with the workflow on Figure 1.



**Fig. 1.** Research workflow.

### 3.2 Method Implementation

In this study, there are two steps are performed. The first step is to choose diagnoses that has a lot of influence by implementing SBFS. There were 18 diagnoses used in the study as shown in Table 1.

**Table 1.** CKD dataset description.

No. Feature	Feature	Type	Value
1	Age	Numerical	Years
2	Blood Pressure	Numerical	mm/Hg
3	Anemia	Nominal	Yes, no
4	Nausea	Nominal	Yes, no
5	Blown	Nominal	Yes, no
6	VAS Scale	Numerical	0, 2, 4, 6, 8, 10
7	Edema	Nominal	Yes, no
8	Dehydration	Nominal	Yes, no
9	Pale Skin	Nominal	Yes, no
10	Weight	Numerical	Kg
11	Pulse	Nominal	Regular, irregular
12	Motion	Nominal	Normal, weak, disturbed
13	Appetite	Nominal	Normal, abnormal
14	Level of Consciousness	Nominal	Compos mentis, apathy, delirium, somnolent, sopor, coma
15	Temperature	Numerical	°C
16	Gender	Nominal	Male, female
17	Another Disease	Nominal	Yes, no
18	Falls (within 3 months)	Nominal	Yes, no

The dataset is divided into 2 classes, CKD class and CKD-A class (acute CKD), with 50 data. The SBFS method eliminates diagnostic features that are considered to have little influence compared to other diagnostic features. Using this method, the elimination is done

backwards by using the initial input of all the features used in the dataset. This method produces the features of diagnosis, which are considered as the best features.

In the second step in the study, the features considered the best diagnoses are used as input to the ANN classification algorithm. The final result of the study shows the best features of the 18 existing features and the accuracy results based on the ANN classification algorithm.

## 4 Result and Discussion

### 4.1 Result

Of the 18 features used in the dataset, based on the implementation of SBFS, one by one, the feature combinations were checked for their effects from one another. Features that have less influence on other features will be automatically eliminated, leaving only the selected features considered the best. The best feature results obtained are as in Table 2.

**Table 2.** The best diagnoses feature.

No. Feature	Feature	Value
1	Age	Years
2	Blood Pressure	mm/Hg
3	Anemia	Yes, no
4	Nausea	Yes, no
5	Blown	Yes, no
6	VAS Scale	0, 2, 4, 6, 8, 10
7	Edema	Yes, no
8	Dehydration	Yes, no
10	Weight	Kg
11	Pulse	Regular, irregular
12	Motion	Normal, weak, disturbed
14	Level of Consciousness	Compos mentis, apathy, delirium, somnolent, sopor, coma
15	Temperature	°C
16	Gender	Male, female
18	Falls (within 3 months)	Yes, no

Based on the 18 diagnostic features used as input in finding the best diagnostic features using SBFS, the remaining 15 features are considered the best by the system. These features are selected based on the combination of features that have the highest percentage of accuracy. While 3 other features of "pale", "appetite", and "another disease" were eliminated automatically in the process because it is considered less influential when compared with other diagnostic features. With a combination of 15 diagnostic features, an accuracy of 86% is obtained.

Furthermore, 15 selected diagnostic features are used as input to the ANN classification algorithm. The distribution of the dataset carried out in this algorithm uses division with details that 70% is used as training data, 15% is used as validation data, and 15% is used as testing data. The accuracy results obtained from the system by implementing the SBFS before ANN are compared with only using ANN, as shown in Table 3.

**Table 3.** Comparison of accuracy.

Condition	Accuracy result
Without using SBFS	80%
Using SBFS	88%

Calculations using evaluation parameters can be seen in Table 4 for processes that only use ANN with all features used as inputs.

**Table 4.** Evaluation parameters without SBFS.

	Predicted CKD	Predicted CKD-A	Recall	Precision
Actually CKD	29	4	0,879	0,829
Actually CKD-A	6	11		

While Table 5 is the calculation of evaluation parameters for the process by implementing SBFS first to get the best feature to be input on ANN.

**Table 5.** Evaluation parameters with SBFS.

	Predicted CKD	Predicted CKD-A	Recall	Precision
Actually CKD	33	4	0,892	0,943
Actually CKD-A	2	11		

The accuracy results displayed by the system may be affected by the type of data in the dataset feature used in the study. In this study, there are more than half the diagnostic features that have binary values (values 0 and 1), so that it can affect the final accuracy result, which results in accuracy that is not so close to 100%. But overall, the application of SBFS, which is able to choose the best diagnosis before implementing ANN, can provide improved accuracy results during processing in the classification algorithm..

## 4.2 Discussion

The study conducted using SBFS and ANN Backpropagation has its advantages and disadvantages. Based on the results and tables, SBFS has the advantage of choosing the best combination of diagnostic features so that there is no need to use all the features to perform a diagnosis. However, SBFS has disadvantages in that it cannot evaluate the features that have been eliminated previously. So, there may be features that are useful but are eliminated by the system. Using the selected features, the use of SBFS and ANN has better results when compared to using only ANN with 88%.

## 5 Conclusion

Based on the study conducted, it can be concluded that the selection of diagnostic features of chronic kidney disease can be made by implementing the SBFS method. The implementation of the method can select 15 diagnostic features that are considered to influence the 18 diagnostic features used in the study. The selection of this feature is based on a comparison of the combinations of each feature that has the highest level of accuracy. Using 15 selected diagnostic features as inputs from the ANN classification algorithm obtained an accuracy of 88%. These results improved when compared to using the overall diagnostic feature, which

has an accuracy of 80%. This accuracy can be influenced by the type of data in the dataset because, in this study, the type of data used has more than half the diagnostic features with binary values. However, it should be noted about the use of SBFS because the method cannot re-evaluate the diagnostic features that have been eliminated. When compared with conventional methods, this method still requires consideration from the medical staff because it is not a final diagnosis for patients.

## References

- [1] E. P. Balogh, B. T. Miller, and J. R. Ball, *Improving Diagnosis in Health Care* (Washington DC, National Academies Press, 2015)
- [2] N. A. Almansour, H.F. Syed, N.R. Khayat, R.K. Altheeb, R.E. Juri, J. Alhiyafi, S. Alrashed, S.O. Olatunji, *Comput. Biol. Med.*, **109**, 101–111 (2019)
- [3] D. Jain dan V. Singh, *Egypt. Informatics J.*, **19**, 179–189 (2018)
- [4] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, *J. Retail. Consum. Serv.*, **27**, 11–23 (2015)
- [5] U. Stańczyk and L. C. Jain, *Feature Selection for Data and Pattern Recognition* (Springer Berlin Heidelberg, 2015)
- [6] D. T. Larose, *Discovering Knowledge In Data : An Introduction to Data Mining*, **3**, (New Jersey: A John Wiley & Sons, Inc., Publication, 2005)
- [7] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models* (CRC Press, 2019)
- [8] N. Gu, M. Fan, L. Du, and D. Ren, *Neurocomputing*, **161**, 199–209 (2015)
- [9] R. Olawoyin, *Chemosphere*, **161**, 145–150 (2016)
- [10] L. Fausett and L. V Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications* (Prentice-Hall, 1994)