p-ISSN: 2301-5373
e-ISSN: 2654-5101

Effect of Feature Selection on Classification Liver Disease

I Wayan Sawendo Eko Wijana[a1], I Gede Santi Astawa[a2], AAIN Eka Karyawati[a3]

[a]Informatics Department, Faculty of Math and Sciences, Udayana University
Badung, Bali, Indonesia
[1]ekosawendo@gmail.com
[2]santiastawa@gmail.com
[3]eka.karyawati@unud.ac.id

Abstract

Classification is the process of differentiating a set of models into several data classes. There are many methods that can be used for the classification process, one of which is the Artificial Neural Network method. Neural networks are a computational method that mimics biological syafar networks. Artificial condition networks can be used to model complex relationships between input and output to recognize patterns in data [1]. In this study, a test was conducted to determine the effect of feature selection on the classification results. This research was carried out by eliminating uncorrelated data variables and correlated data to determine their effect on the classification results and the computation time obtained.In this study, the results show that the accuracy obtained from eliminating uncorrelated data does not really affect the accuracy results where it only has a decrease of 0.04049%, while the correlated features are more influential on the classification results obtained, where the accuracy increases by 1.64%. and 2.02%. For computation time, feature selection does not really affect the computation time obtained.

Keywords: *Classification, Artificial Neural Network, Liver Disease, Accuracy, Time*.

1. Introduction

Classification is the process of differentiating a set of models into several data classes. The classification process aims to predict the class label for data that does not yet have a class label. There are many methods that can be used for the classification process, one of which is the Artificial Neural Network method. Neural networks are a computational method that mimics biological safeguards. Artificial condition networks can be used to model complex relationships between input and output to recognize patterns in data [1]. In Artificial Neural Networks, there are several parameters which are initialized at the beginning and can be changed such as the number of units in the hidden layer. In 2017 a research was conducted on Implementation of Backpropagation Neural Networks to Diagnose Skin Diseases in Children, in this study they tested the number of hidden neurons in order to obtain the smallest MSE error value from each combination of the number of hidden neurons. The results obtained in this study are the optimal number of hidden neurons as many as 4 neurons [2].

In 2016 Nunik Purwaningsih conducted a research on the application of the multilayer perceptron for the classification of the tannest cowhide types. This study aims to apply the MLP method to classify the tanned cow hides. From the test results obtained classification accuracy level reached 87.83%. The most appropriate type of skin that can be identified is pull up skin with an accuracy rate of 98.75% [3]. In 2016, a study was conducted on the diagnosis of Parkinson's disease based on the combination of data mining algorithm and feature selection by Astuti and Ferinanto, which obtained the best accuracy value of 96.923% with a running time of 0.10 seconds for the classification results of data mining with the CFS feature selection. Meanwhile, the Naive Bayes algorithm gets an accuracy value of 76,923% [4]. Research with other objects in 2018 conducted by Amrin and Satriadi on the Implementation of Artificial Neural Networks with Multilayer Perceptron for Credit Lending Analysis obtained an accuracy of 96.1% with an area under the curva (AUC) value of 0.999. From these results it can be said that the classification is done very well, this is because it has an AUC value between 0.90 - 1.00 [5].

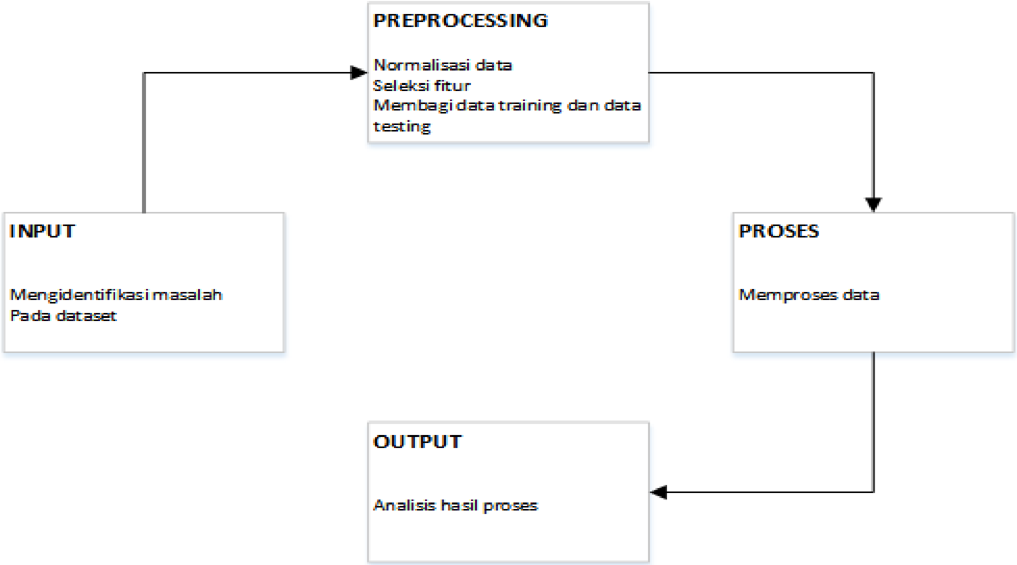
In 2020, research on Influence Optimization Feature Against Liver Disorders Diagnostic Results Using Artificial Neural Network. This study found that data that did not perform feature selection resulted in an accuracy value that tended to be greater than data that did feature selection between 64% and 100%. However, the accuracy value obtained in the data that selected fetuses was more stable than the accuracy values for data that did not select fetuses, namely between 68.57% and 71.42% [6].

In this study, the Artificial Neural Network method will be used for the classification of liver disease. There are many types of liver disease such as hepatitis and liver cancer, but in this study only classified it into a general form, namely positive and negative. A new study by the British Liver Trust reveals that liver disease or liver disease is the leading cause of death in people aged 35-49 years, particularly in the UK. Apart from the liver, at the top of the list of biggest deaths are suicide, heart disease and breast cancer. As reported by netdoctor, this study analyzed data on mortality in England and Wales. The findings show that in 2017, 998 men and women aged between 35 and 49 died from liver disease.

The type of data used in this study is the Liver Disease Lab data obtained from the Kaggle Dataset. This liver disease data has a total of 483 data records. In the process of testing the data will use SPSS for the classification process. In this study, the effect of data attributes on the classification process will be examined, where in the classification process some data items will be removed and their effects on the classification process using the Perceptron Multilayer Neural Network method. In this study, before conducting research on the effect of feature selection, the optimal number of hidden neurons will be sought first so that they will get optimal accuracy.

2. Research Methods

In this research, the research method carried out starts from analyzing the problems to be done, collecting data to be tested, inputting the tested data into the IBM SPSS Statistic 25 tools, seeing the feature correlation to the classification results, creating scenarios from testing, performing the classification process in SPSS, then copy the test results from the classification and computation time. The following is a flow chart of the research to be carried out.



Picture 1. Research Methodology Flow

From the picture above, it can be explained by the following steps:

2.1 Identification of problems

The problem raised is about the selection of features from the most distant or uncorrelated variables to the dependent variable to determine the effect of accuracy and computation time. In addition, this study also examines the effect of changing the number of hidden layer units in Artificial Neural Networks on the results of liver disease classification.

2.2 Data collection

The type of data used in this study is the Liver Disease Lab data obtained from the Kaggle Dataset. This liver disease data has a total of 483 data records.

Table 1. Indian Liver Patient Dataset

Attribute	Description
Age	Patient Age
TB	Total Bilirubin
DB	Direct Bilirubin
Alkphos	Alkaline Phosphotase
Really	Amartotransferse
Sgot	Aspartate
ALB	Albumin
A / G Ratio	Albumin and Globullin Ratio
Class	Dividing Data Into Two Classes 0 and 1

2.3 Data Normalization

Based on the data above, it can be seen that the domain of each feature can be said to be unbalanced, therefore data normalization is needed to equalize the data for each feature so that no feature that has a large value dominates the results of the classification. Here is the formula for data normalization:

$$y = \frac{x - D_{min}}{D_{max} - D_{min}} \quad (1)$$

Information :

y = Value of data after normalization

x = Data values before normalization

D_{max} = The maximum value of all original data

D_{min} = Minimum value of all original data

The dataset will be normalized using formula (1). The data normalization process is done using Microsoft Excel tools.

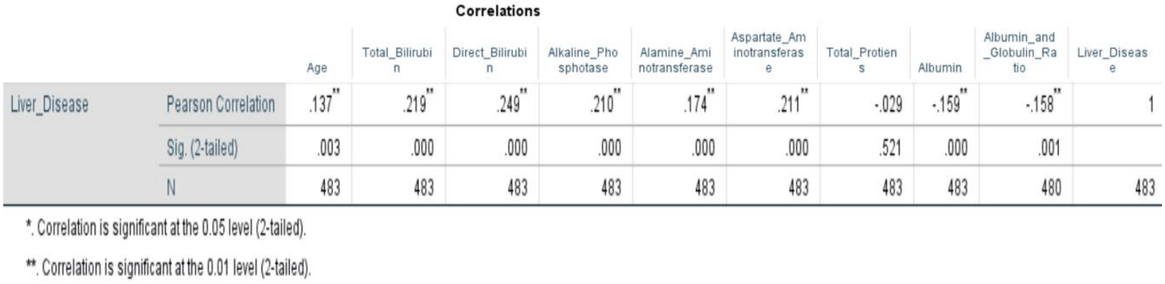
2.4 Perceptron Multilayer Artificial Neural Network

Artificial neural network is one of the artificial representations of the human brain which always tries to simulate the learning process in the human brain. The artificial term here is used because this neural network is implemented using a computer program that is able to complete a number of calculation processes during the learning process [1].

Neural network is a network of a group of small processing units modeled on the basis of human neural networks. Artificial neural networks are adaptive systems that can change their structure to solve based on external information flowing through the network. In simple terms, neural networks are a non-linear statistical data modeling tool. Artificial conditional networks can be used to model complex relationships between input and output to identify patterns in data. Network with multiple layers(multi layer network)It has certain characteristics, namely having 3 types of layers, namely the input layer, the output layer, and the hidden layer. Networks with many layers can solve more complex problems than a network with a single layer. However, the training process often takes a longer time [1].

2.5 Feature Selection

Feature selection is a preprocessing stage that aims to find the value of the relevance of an attribute to the class label and ignore attributes that do not contribute anything to data classification [6]. This feature selection will analyze the significance value of each feature in the dataset used in the study. Where if the significance value of a feature is close to 0 then the feature has a strong influence on the output dataset. And if the feature significance value is close to 1, the feature will not have a significant effect on the dataset output. Before performing feature selection, a significant value search will be carried out for each feature to the output using the Bivariate Correlation feature on IBM SPSS. The results of the Correlation analysis can be seen in the following Figure:



Correlations

		Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Liver_Disease
Liver_Disease	Pearson Correlation	.137**	.219**	.249**	.210**	.174**	.211**	-.029	-.159**	-.158**	1
	Sig. (2-tailed)	.003	.000	.000	.000	.000	.000	.521	.000	.001	
	N	483	483	483	483	483	483	483	483	480	483

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 2. Correlation Testing Results

Where if the significance value of a feature is below or equal to 0.05, this feature has a strong influence on the output dataset. And if the feature significance value is more than 0.05, the feature will not have a significant effect on the dataset output. After checking the correlation (can be seen in Figure 2), the results show that the total protein has a correlation value above 0.05 and does not really have an effect on the output dataset, so that in later testing, the total protein will be removed to determine its effect on the test results. In addition to total protein, age as well as albumin will be removed because it has a correlation value below 0.05 and affects the output dataset.From these results in this study several scenarios for selecting features / attributes used in data processing will be determined as follows:

Table 2. Testing Scenarios

Scenario	Selected Features
1	All Data
2	Without total protein
3	Without Age
4	Without Albumin

Scenario 1 is done in order to find out how much accuracy will be obtained if all the data is used and also as a comparison to other scenarios. Scenario 2 is carried out in order to determine the effect of uncorrelated data variables on the classification results obtained. Scenarios 3 and 4 are carried out in order to determine the effect of correlated data on the classification results obtained.

2.6 Data Processing

The classification method used in this research is the Artificial Neural Network Multilayer Perceptron method. This is done by finding the optimal number of hidden layer neurons by comparing the number of hidden layers used. The maximum duration of training used is 15 minutes, the maximum epoch given is 1000 and the proportion of training data and test data used is 70:30. After obtaining the optimal number of hidden neurons, the effect of feature selection on the classification of liver disease will be carried out using the number of hidden neurons that have been obtained and using the maximum training time, maximum epoch, the same data proportion as when testing the number of hidden layer neurons.

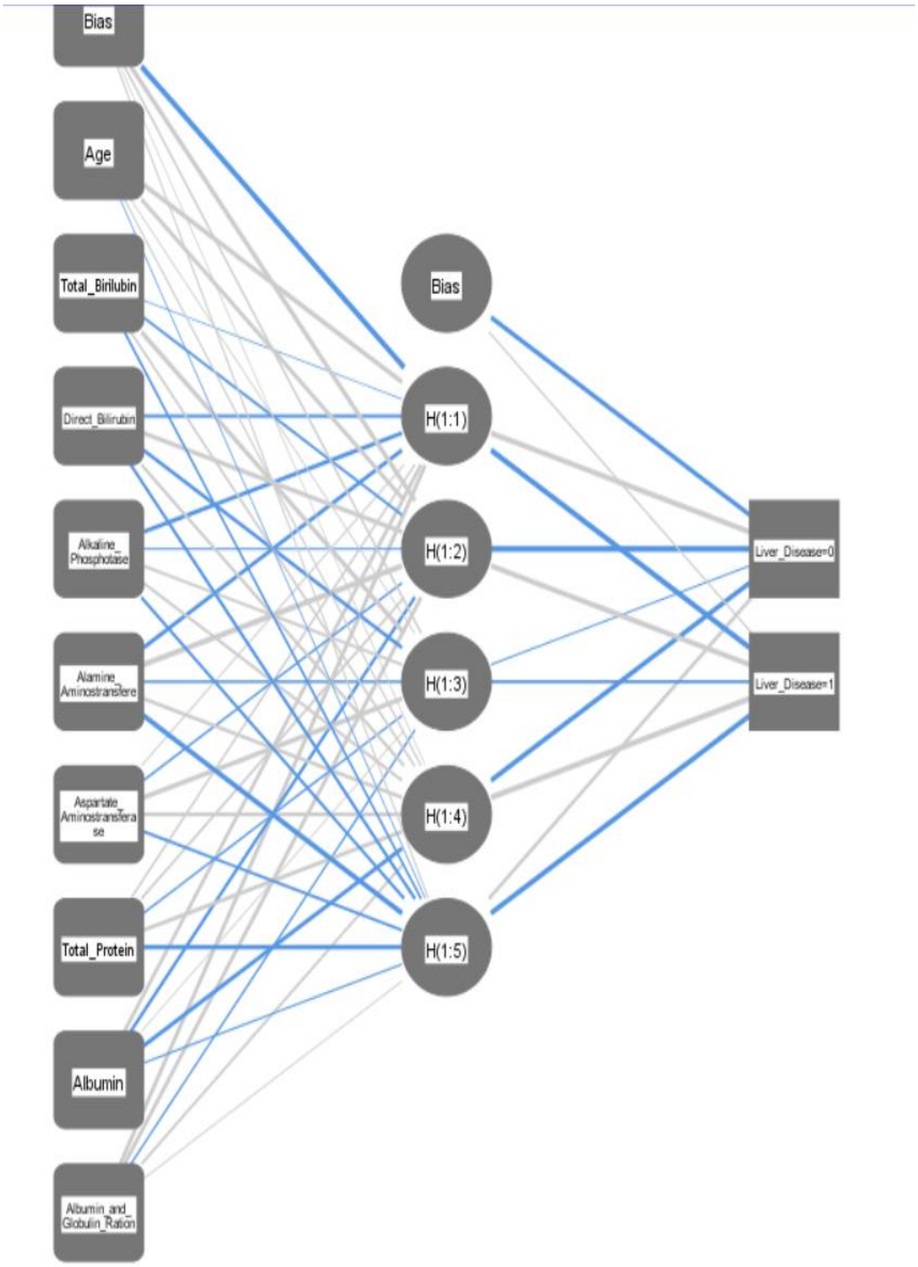


Figure 3. Example of ANN Architecture

2.7 Analysis of Data Processing Results

The things that will be analyzed in this study are the effect of reducing correlated and uncorrelated features of the target on the accuracy and computation time of the classification of liver disease using the Multilayer Perceptron method using SPSS.

3. Results and Discussion

3.1 Data Processing

At the data processing stage, a Perceptron Multilayer Neural Network will be implemented which will be carried out on the IBM SPSS tool. In IBM SPSS, there is a Neural Network feature that can be used to apply the Multilayer Perceptron method. The implementation used is the same as that described at the data sharing stage, namely 7: 3.

1. **Testing the number of hidden neurons.**

In testing the number of hidden neurons, the aim is to obtain the optimal number of hidden neurons so as to increase the accuracy obtained. The results obtained can be seen in the table below:

Table 3. Hidden Layer Measurement Results

Number of Units on Hidden Layer	Accuracy	Time
2	71.2	0.09
4	72.5	0.1
6	71.8	0.08
8	72.7	0.1
10	73.8	0.11
12	73.1	0.12
14	72.9	0.1
16	73.9	0.1
18	73.8	0.11
20	74.2	0.12
22	74.12	0.1
24	74.05	0.13

From the experiments that have been carried out, the optimal number of hidden neurons is as many as 20 with an accuracy obtained of 74.2%, but the number of hidden neuron units is also used to find the effect of feature selection

2. **The effect of feature selection**
   The effect of feature selection to determine the accuracy and computation time of the classification will use the number of hidden neurons that have been obtained, namely as many as 20 with a maximum training time of 15 minutes. The results of accuracy and computation time can be seen as follows:

Table 4. Feature Selection Results

Trial	All Data		Without total protein		No Age		Without Albumin	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
1	72.1	0.12	72.5	0.15	76.7	12	75.9	0.11
2	74.5	0.14	73.4	0.1	76.6	0.16	74.2	0.14
3	73.7	0.17	72.9	0.13	77.6	0.19	74.7	0.13
4	74.7	0.11	78.6	0.1	75.3	0.12	76.6	0.1
5	74.8	0.16	74.6	0.15	75.3	0.11	75.7	0.14
6	74.5	0.14	73.9	0.13	72.5	0.16	75.5	0.14
7	75.3	0.14	74.7	0.13	75.6	0.13	76.6	0.13
8	75.3	0.11	73.5	0.13	76	0.18	76.2	0.12

9	73.7	0.14	72.1	0.14	74.5	0.17	75.9	0.11
10	72.2	0.17	74.3	0.13	72.9	0.14	74.5	0.12
Average	**74.08**	**0.14**	**74.05**	**0.129**	**75.3**	**1,336**	**75.58**	**0.124**

3.2 Results Analysis

The results of the Multilayer Perceptron implementation which aims to determine the effect of feature selection on the classification results shown in Table 4 show that the feature selection results do not really affect the computation time. For the accuracy results obtained, uncorrelated data does not really affect the results of accuracy (has a small difference when compared to scenario 1), where the accuracy obtained decreases by 0.04% and correlated data has more influence on the classification results, which can be where the accuracy increased by 1.64% and 2.02%, it can be seen from the following table:

Table 5. Percentage Decreased Computing Time

Scenario	Average Accuracy	Difference (with scenario 1 accuracy)
1	74.08	0%
2	74.05	0.04049% decrease
3	75.3	1.64% increase
4	75.58	2.02% increase

4 Conclusion

This study aims to determine the effect of using feature selection and without using feature selection for classification of liver disease and also to examine the effect of changing the number of units in the hidden layer on the results of classification of liver disease. The type of data used in this study is the Liver Disease Lab data obtained from the Kaggle Dataset. This liver disease data has a total of 483 data records. Where in comparing the results of using feature selection, it is divided into 4 scenarios that have been determined by searching for the significance value with the SPSS correlation.

In this study, the results show that the accuracy obtained from eliminating uncorrelated data does not really affect the accuracy results where it only has a decrease of 0.04049%, while the correlated features are more influential on the classification results obtained, where the accuracy increases by 1.64%. and 2.02%.For computation time, feature selection does not really affect the computation time obtained. In the future, it is hoped that more research on feature selection will be carried out using different data and changes in different ANN parameters so that it will get better accuracy.

Reference

[1] Solikhun, and M. Wahyudi, Jaringan Syaraf Tiruan Backpropagation Pengenalan Pola Calon Debiur, Medan, Yayasan Kita Menulis, 2020.

[2] R. S. Suhartanto, C. Dewi, and L. Muflikhah, “Implementasi Jaringan Syaraf Tiruan Backpropagation untuk Mendiagnosis Penyakit Kulit pada Anak” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol.1, no. 7, 2017.

[3] Nunik Purwaningsih, "Penerapan Multilayer Perceptron Untuk Klasifikasi Jenis Kulit Sapi Tersamak" Jurnal TEKNOIF, vol. 4,no 1, 2016.

[4] T. Astuti and T. Ferinanto, "Diagnosis Penyakit Parkinson Berdasarkan Kombinasi Algoritme Data Mining Dan Seleksi Fitur," Seminar Nasional APTIKOM (SEMNASTIKOM), pp. 127-130, 2016

[5] A. and I. Satriadi, "Implementasi Jaringan Syaraf Tiruan Dengan Multilayer Perceptron Untuk Analisa Pemberian Kredit," Jurnal Riset Komputer (JURIKOM), vol. 5, pp. 605-610, 2018.

[6] K. D. Prebiana and I G. S. Astawa, “Influence Optimization Feature Against Liver Disorders Diagnostic Results Using Artificial Neural Network”, Jurnal Elektronik Ilmu Komputer Udayana, vol. 8, no.3, 2020.