Classification of Heart Disease Using Linear Discriminant Analysis Algorithm

R. Rizal Isnanto¹, Ibnu Rashad², and Catur Edi Widodo³

¹Computer Engineering Department, Faculty of Engineering, Diponegoro University
²Master of Information Systems Program, Postgraduate School of Diponegoro University
³Physics Department, Faculty of Science and Mathematics, Diponegoro University

Abstract. Ischaemic coronary heart disease is the number one cause of death globally. Detecting this disease can only be done by consulting directly with a cardiologist at a cost that is certainly not small. Therefore, is a need for a system to detect heart disease in patients with accuracy but low cost. With the development of technology, especially in artificial intelligence area, there was machine learning techniques to enhance automatic detection capabilities. Linear Discriminant Analysis are one of machine learning method for prediction to detect heart disease as early as possible. In this study, linear discriminant analysis algorithm was implemented to classify heart disease. Dataset used are from the UCI machine learning repository. This study carried out two experimental conditions, classifying heart disease based on suffer or not, other is classifying heart disease by 5 level stage. Result proves that the performance of the classifier with LDA with 2 classes is better than 5 classes. Performance of the LDA algorithm in classifying heart disease with 2 labels that are used as targets or outputs. From these results, the precision value is 0.82, the recall value is 0.81, the F1 score value is 0.81, with an accuracy of 81.22%.

1 Introduction

Heart disease is a general term for all disorders that originate from the heart. Cardiovascular disease (CDV) is the main cause of death in the world compared to other diseases. Heart disease is a non-communicable disease that causes the most deaths. Based on the report of the World Health Organization (WHO) in 2000, which can make a person unable to recognize the initial symptoms so that initial treatment becomes difficult. Currently, ischemic coronary heart disease is the first cause of death globally. The average heart disease tends to increase from year to year. It is predicted that in the coming years it will cause more deaths [1].

There is still a lack of awareness of healthy lifestyles and a lack of information about coronary heart disease. The process for detecting heart disease can be done manually, namely by consulting directly with a cardiologist and bringing some test results back to the

* Corresponding author: rizal_isnanto@yahoo.com
laboratory, followed by consulting back to the cardiologist. This step is relatively expensive. With a large risk of death, we need a system that can detect heart disease in patients that is accurate and low cost [2].

With the development of technology and information in the field of artificial intelligence, the introduction of machine learning techniques helps improve automatic detection. The methods often used in machine learning are Linear Discriminant Analysis and Linear Discriminant Quadratic prediction so that heart disease can be detected as early as possible. With the help of a detection system, the possibility of misdiagnosis by experts can be avoided, and medical data can be checked as early as possible. And medical data can be checked in a short span of time and is more precise and accurate [3].

Various prediction studies have been carried out to determine the level of accuracy produced by each method and the influencing factors. Research related to the application of the prediction method of linear discriminant analysis was conducted by Sulistio. The linear discriminant analysis method is used in the prediction module which separates groups of students into their respective classes (graduated on time or late) and then shows the results to the user. The data used for input is the subject value of each student. The subjects used in the application were selected by the head of the department with 22 subjects selected, for evaluation purposes the cross-validation method was used with an average accuracy of 97% [4].

Previous research on the accuracy of the linear discriminant analysis method was refined by Budiman's research entitled Detection of Types of Autism in Early Childhood Using Linear Discriminant Analysis. His research used training data from 75% of the data used and the remaining 25% as test data to test the system model resulting from the use, types of autism in early childhood. From the results of his research the linear discriminant analysis method shows an accuracy of 88%, but the accuracy can change when training data is added to the dataset [5].

Prediction methods that are often used include Logistic Regression (LR) algorithms, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis, and K-Nearest Neighbor (KNN). In this research, the formulation of the problem in developing a classification method is using LDA (Linear Discriminant Analysis) for cases of heart disease in its application directed at the high-risk category (for sufferers) to the low-risk category (not). The linear discriminant analysis method is used for pattern recognition and machine learning in finding linear combinations that characterize or separate two or more objects or events. The combination obtained can be used as a linear classification or commonly used as a dimension reduction process before classifying, so that values with more accurate and precise accuracy can be produced.

2 Basic Theories of Data and Classification

2.1 Data Mining

Data mining is a term used to describe the discovery of knowledge in databases. The term data mining has been widely used by statisticians, data analysts and the Management Information Systems (MIS) community. Data mining is the automatic analysis of large or complex data with the aim of finding important patterns or trends that are usually not realized. [6].

Data mining can discover hidden trends and patterns that do not appear in simple queries and so can have an important part in finding knowledge and making decisions. As a technological tool, data mining can be applied to all types of data as long as the data is meaningful for building an application. The most basic forms of data to support the basic methods of data mining are databases, data warehouses, transactional data. Data mining can also be applied to other forms of data such as data streams, graphic or network data, spatial
data, text data, multimedia data and others. The outline of the data mining method can simplify the work process of the system in processing various forms of text [7].

2.2 Min-Max Scalar

Min-Max Scalar works by scaling data or adjusting data within a certain range (range of minimum values to maximum values), with the range usually used from 0 to 1. The following is a mathematical description.

\[ X_{sc} = \frac{x - x_{min}}{x_{max} - x_{min}} \]  

(1)

Where \( x \) is the original value, \( x_{sc} \) is the normalized value. Let's say we have a patient data with the patient's weight, and the patient's body range (100 pounds, 150 pounds). To rescaling this data, we first subtract 100 from each patient's weight and divide the result by 50 (the difference between the maximum and minimum weights). To change the scalar range between changing values – change \([a, b]\) the formula to be:

\[ X = a + \frac{(x - x_{min})(b - a)}{x_{max} - x_{min}} \]  

(2)

Where \( a, b \) is min-max value.

2.3 Classification Algorithm

Classification Algorithm is an algorithm that is widely used for purposes of classification in the field of machine learning. Classification algorithms are usually used as a calculation method to predict the characteristics of new data with old data. Classification algorithms also usually refer to mathematical formulas or functions that also have predictive abilities by classifying new data into certain classes. Some examples of classification algorithms are linear classifiers, support vector machines, quadratic classifiers, kernel estimation, boosting, decision trees, neural networks and vector quantization learning. All classification algorithms have their advantages and disadvantages, depending on various factors such as the type of data set, the number of classes, etc. [8].

2.4 Linear Classification

In statistical classification, the purpose or objective of classification is to determine or identify where an object belongs to a particular group or class. This is done by using a linear combination of several object properties or characteristics which are processed by the linear prediction function for each available class. Then, based on the evaluation results, you can determine that the new object is in a certain class by passing the new object's properties in the predictor. The function value obtained from this classification process determines the object class after it is interpreted. One of the advantages of linear classification is the speed of classification, especially for objects that have only two classes. Linear classification can distinguish between identification models and identification functions which are training methods and identification models to interpret the results of identification models and identification functions. An example of the algorithm used for the identification model is the linear identification analysis method [9].
2.5 Linear Discriminant Analysis

Linear discriminant analysis, abbreviated as LDA, is a generalization of the linear discriminant angle. It is a method used in statistics, pattern recognition, and machine learning to find combinations of linear features, or to separate two or more objects or events. The resulting combinations can be used as linear classifiers or commonly used in initial dimension reduction. The purpose of linear discriminant analysis (LDA) is to classify objects into a number of classes according to the characteristics described. In linear discriminant analysis, an object has two variables: related class/variable (the dependent variable) and unrelated attribute/variable (independent variable). The dependent variable is associated with the independent variable that describes the variable [10].

These independent variables will later be used to determine the linear combination of these objects. LDA works by using dispersion matrix analysis to find the optimal projection so that it can project the input data into a smaller space where all samples can be separated widely. In LDA, the dependent variable is a class of object which usually has a nominal value/class name and the independent variable is a feature that describes the object which is usually a scalar value. Before making predictions, LDA requires a learning step to determine the discriminant function. This training step requires objects that have been classified with several independent characteristics/variables. Steps to calculate training steps in LDA i.e. [10]:

1. Group the training data into a matrix of a number of classes denoted by Xi, where i = number of classes
2. The next step is to calculate the average matrix for each class (μi).
3. Compute the global average of all data matrices (μ),
4. After getting the following global average we have to calculate the average of the corrected data Xi by subtracting each Xi value from the global average value (μ).
5. Calculate the covariance matrix of each (C) of each Xi with the formula:

\[
C_i = \frac{(x_i^0)^T x_i^0}{n_i}
\]

Description:
\(n_i = \) the number of rows of group Xi

6. Calculate the value of the global covariance matrix (C) by using the formula:

\[
C = \frac{1}{N} \sum_{i=1}^{g} n_i C_i
\]

Description:
\(N = \) the number of rows of the total data

7. Calculate the reciprocal value of the global covariance matrix (C^−1),
8. After getting the matrix inverse, it is possible to calculate each class (Pi).
9. That way we can calculate the discriminant function (fi), with the formula:

\[
f_i = \mu_i C^{-1} x_k^T - \frac{1}{2} \mu_i C^{-1} \mu_i^T + \ln (p_i)
\]

After each class is known, the accuracy of the discriminant function (fi) can be determined by re-mapping each group in the training data with the discriminant function fi. The training data object will be included in the class depending on the maximum value. After the discriminant formula is known and tested for accuracy, the training phase can be declared.
complete, and this discriminant function can be used to classify new objects into the specified class by calculating the $X^T_k$ of the new object and then plugging it into the discriminant function of each class. The new object will be grouped into the class represented by the owning discriminant function.

### 2.6 Confusion Matrix

The Confusion Matrix is a commonly used method for performing accurate calculations on data mining concepts. Below is an example of a confusion matrix in a classification with two classes [11]. Figure 1 depicts the confusion matrix.

![Confusion Matrix Diagram](image)

**Fig. 1.** Confusion Matrix

In the two-class confusion matrix, the numbers on the diagonal from left to right are the correct prediction results, and the numbers outside the diagonal are the wrong prediction results.

Measuring the performance of a classification system is very important [12]. System performance classification shows how good the system classifies data. The confusion matrix contains information that compares the system classification results with the actual classification results [13]. Performance analysis used in this study includes Accuracy (6), Precision (7), Recall (8), and F1 (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP+TN}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F_1 = \frac{2TP}{2TP+FP+FN} \quad (8)$$

### 2.7 Correlation Coefficient

The correlation coefficient is the value used to measure the relationship between two pairs of data sets. The value of this correlation can be categorized as follows [14].
1. Positive correlation: if one variable increases, then the other variables will also increase, and vice versa.
2. Negative correlation: if one variable increases, the other variable decreases, and vice versa.
3. No Correlation: when an increasing in variable does not affect other variables at all.
4. Positive correlations can be grouped again according to the number of values as follows:
   a. -1.00: perfect negative linear relationship
   b. -0.99 until -0.70: strong negative linear relationship
   c. -0.69 until -0.50: medium negative linear relationship
   d. -0.49 until -0.01: weak negative linear relationship
   e. 0: there is no linear relation.
   f. 0.01 until 0.49: weak positive linear relationship
   g. 0.50 until 0.69: moderate positive linear relationship
   h. 0.70 until 0.99: strong positive linear relationship
   i. 1.0: perfect positive linear relation.

There are 10 (ten) steps to calculate the correlation coefficient as listed below.
1. Calculate the total row column x or column y \(n\)
2. Calculate the total of the squared values in column x \(\Sigma x^2\)
3. Calculate the total of the squared values in column y \(\Sigma y^2\)
4. Calculate square of the total value in column x \(\Sigma x^2\)
5. Calculate square of the total value in column y \(\Sigma y^2\)
6. Calculate the total of the value x multiplied by y \(\Sigma xy\)
7. Calculate \(S_{xx} = \Sigma x^2 - (\Sigma x)^2\)
8. Calculate \(S_{yy} = \Sigma y^2 - (\Sigma y)^2\)
9. Calculate \(S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)\)
10. Calculate correlation coefficient by formula \(p = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\)

3 Research Methodology

The classification system was built using the Linear discriminant analysis algorithm in predicting heart disease. The stages of the flow of this research method can be seen from Figure 2.

In this research, the classification method used to classify datasets is linear discriminant analysis. The dataset used is divided into 25% test and 75% training. This classification is carried out under two conditions. The first condition is the number of outputs consisting of 5 labels and the second condition is only the number of labels with 2 outputs. The classification of performance measurement based on accuracy, precision, repeatability, and F1 value.
4 Results and Discussion

This research uses the dataset shown in Table 1. The dataset is originated from the link of https://archive.ics.uci.edu/ml/datasets/heart+disease. The dataset used in this study has nine attributes as input values for the classification of heart disease. There are two conditions for using LDA when carrying out this classification. The first is the output value for 2 labels, then the output value for 5 labels.

Table 1. Dataset of research.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>1</td>
<td>4</td>
<td>140</td>
<td>239</td>
<td>0</td>
<td>0</td>
<td>160</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0</td>
<td>3</td>
<td>130</td>
<td>275</td>
<td>0</td>
<td>0</td>
<td>139</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td>2</td>
<td>130</td>
<td>266</td>
<td>0</td>
<td>0</td>
<td>171</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>1</td>
<td>1</td>
<td>110</td>
<td>211</td>
<td>0</td>
<td>2</td>
<td>144</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>4</td>
<td>140</td>
<td>266</td>
<td>0</td>
<td>0</td>
<td>134</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>4</td>
<td>160</td>
<td>331</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>0</td>
<td>3</td>
<td>130</td>
<td>294</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>1</td>
<td>4</td>
<td>155</td>
<td>342</td>
<td>1</td>
<td>0</td>
<td>150</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td>4</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>126</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>4</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>120</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>4</td>
<td>135</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>128</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>4</td>
<td>165</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>122</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td>4</td>
<td>110</td>
<td>206</td>
<td>0</td>
<td>2</td>
<td>108</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td>3</td>
<td>130</td>
<td>228</td>
<td>0</td>
<td>1</td>
<td>140</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td>4</td>
<td>130</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>140</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>1</td>
<td>4</td>
<td>124</td>
<td>171</td>
<td>0</td>
<td>1</td>
<td>110</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>1</td>
<td>4</td>
<td>136</td>
<td>248</td>
<td>0</td>
<td>0</td>
<td>140</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the results of the performance of the LDA algorithm in classifying heart disease using the two labels used as targets or results. Based on these results, the precision value is 0.82, the repetition value is 0.81, the F1 value is 0.81, and with an accuracy of 81.22%. Figure 3 shows the confusion matrix which is found in classic heart disease with LDA at 2 targets or outputs.
Table 2. LDA performance results with 2 target labels.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Disease</td>
<td>0.85</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>0.78</td>
<td>0.87</td>
<td>0.82</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Accuracy = 81.22%

Fig. 3. Confusion Matrix results from LDA with 2 target labels

While, Table 3 indicates the performance result of the LDA algorithm in classifying heart disease with 5 labels used as targets or outputs. Based on these results, the precision value is 0.56, the repetition value is 0.59, the F1 value is 0.56, with an accuracy of 59.38%. Figure 4 depicts the confusion matrix which is found in classic heart disease with LDA at 5 targets or outputs.

Table 3. LDA performance results on 5 target labels

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Heart Disease</td>
<td>0.76</td>
<td>0.85</td>
<td>0.80</td>
</tr>
<tr>
<td>Heart Disease stage 1</td>
<td>0.46</td>
<td>0.60</td>
<td>0.52</td>
</tr>
<tr>
<td>Heart Disease stage 2</td>
<td>0.26</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>Heart Disease stage 3</td>
<td>0.26</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td>Heart Disease stage 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>0.56</td>
<td>0.59</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Accuracy = 59.38%

This study conducted tests using Linear Discriminant Analysis to classify heart disease. This is also done with two output value conditions, the first with 2 targets or labels and the second with 5 targets or labels. The comparison purpose is to analyze the performance of the linear discriminant analysis algorithm in classifying it.
Fig. 4. Results of LDA Confusion Matrix with 5 target labels

Based on Figures 5, 6, 7 and 8 it can be seen that the performance of LDA with class 2 output is better than class 5 output. This happens because the amount of input data in 5 classes is not balanced, this can be seen in Figure 5 which is shown in the matrix confusion.

Fig. 5. Comparison of LDA accuracy based on target class

Fig. 6. Comparison of LDA precision by target class
5 Conclusions

Based on the research results, we can conclude from this study that the classifier performance with 2 LDA classes is better than 5 classes. The performance of the LDA algorithm in classifying heart disease using a target or two labels is used as the result. Based on these results, the accuracy value was 0.82, the repetition value was 0.81, the F1 score was 0.81, and the accuracy was 81.22%. The results of the performance of the LDA algorithm in the classification of heart disease using 5 markers are used as targets or outputs. Based on these results, the accuracy value obtained was 0.56, the repetition value was 0.59, and the F1 score of 0.56 was 59.38%. This research is still in the development stage by comparing various other classification methods.

References


7. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Elsevier, 3rd edition, United States of America (2012)


