BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0197–0210
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0197-0210

COMPARISON OF ANN METHOD AND LOGISTIC REGRESSION METHOD ON SINGLE NUCLEOTIDE POLYMORPHISM GENETIC DATA

Adi Setiawan[1*], Rachel Wulan Nirmalasari Wijaya[2]

[1,2]*Department of Mathematics and Data Science, Faculty of Science and Mathematics, Universitas Kristen Satya Wacana*
Salatiga, 50711, Indonesia

*Corresponding author's e-mail: * **adi.setiawan@uksw.edu***

Article History:
Received: 20th September 2022
Revised: 13th December 2022
Accepted: 24th January 2023

Keywords:
Classification;
Neural Network;
Single Nucleotide Polymorphism.

ABSTRACT

This study aims to determine the goodness of classification using the ANN method on Asthma genetic data in the R program package, namely SNPassoc. SNP genetic data was transformed using codominant genetic traits, namely for genetic data AA, AC, CC were given a score of 0, 0.5 and 1, respectively, while CC, CT and TT were scored 0, 0.5 and 1, respectively. The scoring is based on the smallest alphabetical order given a low score. The average accuracy, precision, recall and F_1 score were determined using the neural network method if the genetic code was used with variations in the proportion of test data 10%, 20%, 30% and 40% and repeated B = 1000 times. The results obtained were compared with the logistic regression method. If 20% test data is used and the ANN method is used, the accuracy, precision, recall and F_1 scores are 0.7756, 0.7844, 0.9844 and 0.8728, respectively. When all information from various countries is used in the Asthma genetic data, the logistic regression method gives higher average accuracy, precision and F1 scores than the ANN method, but the average recall is the opposite. When a separate analysis is performed for each country, the logistic regression method gives higher accuracy, precision, recall and F_1 scores in the ANN method compared to the logistic regression method.

How to cite this article:

A. Setiawan and R. W. N. Wijaya., "COMPARISON OF ANN METHOD AND LOGISTIC REGRESSION METHOD ON SINGLE NUCLEOTIDE POLYMORPHISM GENETIC DATA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0197-0210, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

The human genome set consists of the genetic code, namely adenine (A), guanine (G), thymine (T), and cytosine (C). This genetic code consists of about 3×10^9 sequences A, G, T, or C found on human chromosomes consisting of 23 pairs. Therefore, the genetic code at the same location is paired. Single Nucleotide Polymorphism (SNP) is a difference in the composition of a single nucleotide base in the genome of an individual that causes genetic variation in a population. SNPs can be used as markers to find out whether there is a link between the genes present at the SNP location and certain diseases of concern. It can also be done by seeking answers to whether there is a relationship between certain traits/diseases and certain SNP locations. There are many in-depth and recent studies on the association between SNP (Single Nucleotide Polymorphism) markers and certain diseases [1]–[4].

Classification can be used in the analysis of SNP data, namely by classifying each individual's SNP data to determine whether a particular individual is likely to be associated with a particular disease of concern or not. Likewise, whether the SNP pool is also associated with a particular trait/disease. Various classification methods can be used, including the KNN (*k*-nearest neighbor) method, the nave Bayes method, the RF (random forest) method, the SVM (Support Vector Machine) method, the logistic regression method, and the neural network method. Research related to the classification of SNP data are [5]–[7]; however, there has not been much comparison of the goodness between these methods. In this research, it will be conducted on the comparison of the logistic regression method and the neural network method in the classification of SNP data.

Regression can be used in the analysis of SNP data by performing a simple regression analysis (only one SNP is used) or several SNPs used in the model and associated with certain response variables [8], [9]. If the response variable is binary data, a logistic regression model can be used.

In this study, the question of which method is better in classifying the class of cases or controls of asthma based on several closely related SNPs will be answered.

2. RESEARCH METHODS

In this research, it is presented about simple linear regression analysis and multiple linear regression analysis, logistic regression method and ANN method.

Simple linear regression analysis is a function that is used to make predictions about one response variable (the dependent variable) based on known information about another variable called the explanatory variable (the independent variable). Multiple linear regression analysis is based on the following assumptions:

- There is a linear relationship between the response variable and the independent variable,
- The independent variables are not highly correlated with each other,
- Observations are selected independently and randomly from the population,
- Residues should be normally distributed with a mean of 0 and constant variance.

The coefficient of determination (R^2) is a statistical measure used to measure how much variation in the response variable can be explained by the variation of the independent variable. R^2 will increase if more independent variables are used in multiple linear regression, even though the independent variables may not be related to the response variables. R^2 can only be between 0 and 1, where 0 indicates that the result cannot be predicted by any of the independent variables and 1 indicates that the result can be predicted without error from the independent variable. In this study, the question of which method is better in classifying the class of cases or controls of asthma based on several closely related SNPs will be answered.

Suppose in simple regression analysis a model is used.

$$y = b_0 + b_1 x \quad (1)$$

where y represents the response variable, x represents the independent variable/predictor, b_0 represents the intercept and b_1 represents the gradient or slope, then in multiple linear regression analysis it is expressed by

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n. \quad (2)$$

Information about simple linear regression analysis and multiple regression analysis is expressed by [10] and [11].

Binary logistic regression is used to predict the probability that an outcome has only two values (*dichotomy*). Prediction is based on the use of one or more predictors (independent variables) that have numerical or categorical values. Linear regression analysis cannot be used to predict this value considering two reasons, namely linear regression analysis will predict values outside the acceptable range, namely outside the value of 0 or 1. In addition, because the response variable only has two possible values, namely 0 or 1 such that the residue is not normally distributed. The binary logistic regression model associated with simple regression analysis is expressed by

$$p = \frac{1}{1 + \exp\left(-(b_0 + b_1\, x)\right)} \tag{3}$$

while in relation to multiple linear regression analysis, the binary logistic regression model is expressed by

$$p = \frac{1}{1 + \exp\left(-(b_0 + b_1\, x_1 + b_2 x_2 + \cdots . + b_n x_n)\right)} \tag{4}$$

where p represents the probability, b_0 represents the intercept, x_i represents the independent variable and b_i represents the coefficient of the independent variable x_i for i = 1, 2, …, n. Using the logit transformation (log-odds) we get

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots . + b_n x_n. \tag{5}$$

In classification problems using genetic data, the response variable is case/control status of asthma, while the explanatory variable is one or several SNPs. Parameter b_i is estimated using MLE (maximum likelihood estimator). More information about this can be seen in **[12]**, **[13]**.

Artificial Neural Network (ANN) is a computational network that attempts to simulate decision making in the neuronal network of a biological (human or animal) central nervous system. This simulation uses biological knowledge so that it is different from conventional computing machines (digital or analog) which function to replace, enhance or accelerate the computation of the human brain regardless of the arrangement of computing elements and networks **[14]**.

A different aspect of ANN that benefits conventional computers is their high parallelism. Conventional digital computers are sequential machines. If one (out of millions) of transistors fails, then the whole machine stops. In the human central nervous system, thousands of neurons die each year, but brain function is completely unaffected, except when cells in very few important locations die in very large numbers (e.g., severe stroke).

The perceptron is the earliest neural computing model created by F. Rosenblatt and originated in 1958. The perceptron has a basic structure as described by a nerve cell (biological neuron) in **Figure 1**, from several weighted input connections connected to the output, several neurons on the side. Input and output cells are connected to some other nerve cells on the output side. Mathematically the perceptron can be depicted in **Figure 2**.

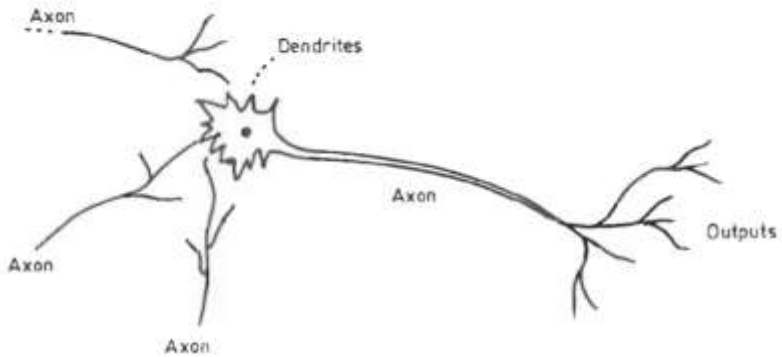

Figure 1. Biological Neuron

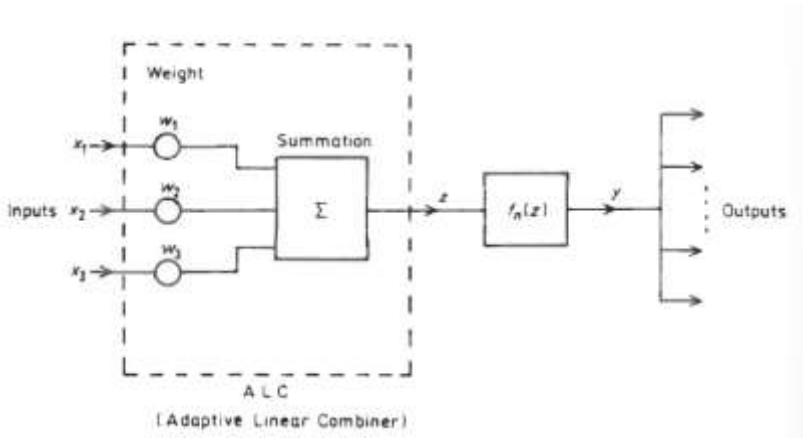

Figure 2. Mathematics Symbol of Perceptron Based on Biological Neuron Idea

As an example, suppose it is denoted z_i as the sum of the outputs of the i-th perceptron and $x_{1i}, \cdots, x_{ni}$ as inputs. The Perceptron cell output differs from the sum of the above equations because of the cell body activation operation, such as the output of a biological cell that is different from the sum of its input weights. The activation operation is an activation function $f(z_i)$ which is a non-linear function that produces the i-th cell output (y_i). The activation function f is also known as squashing function as it keeps the cell output between certain limits as in biological neurons. Various types of functions $f(z_i)$ are used, all of which have limiting properties. The most common activation function is the sigmoid function which is a continuous differentiated function that satisfies the limit. The output and error calculation algorithms are described below:

1. Initialize weight w_{ij} from j-th input to i-th cell and input data $x_{1i}, \cdots, x_{ni}$.
2. It is calculated the number of outputs z_i from the i-th perceptron

$$z_i = \sum_{j=1}^{m} w_{ij} x_{ij}. \quad (6)$$

3. The perceptron cell output is calculated using the activation function, that is, the sigmoid function is selected

$$y_i = \frac{1}{1 + \exp(-z_i)} = f(z_i). \quad (7)$$

4. Steps 2 and 3 are repeated until the output value is obtained from the output layer.
5. Next, look for the error from the training data output compared to the expected value

$$\varepsilon \triangleq \frac{1}{2} \sum_{k} (d_k - y_k)^2 \quad (8)$$

where d_k is the expected output of y_k.

The simplest Perceptron arrangement is the single-layer Perceptron. The single-layer way of working is that the input layer is projected directly to the output layer of the neurons. However, in 1969, Minsky and Papert [15] demonstrated the limitations of the single-layer perceptron. They show that the perceptron cannot even solve a simple Exclusive-OR (XOR) problem.

To overcome these limitations, we need something beyond single-layer ANN. In 1986, Rumelhart, Hinton, and Williams [16] showed that 2-layer ANN could solve the XOR problem above. Extending to three or more layers expands the class of problems that ANNs can solve. However, in the 1960s and 1970s, there were no tools that could be used to set up multi-layer ANNs. In 1986 the backpropagation (BP) algorithm was introduced by Rumelhart, Hinton, and Williams [16] to assign weights and for training multi-layer perceptrons. The BP algorithm starts by calculating the output layer, which is the only one where the desired output is available.

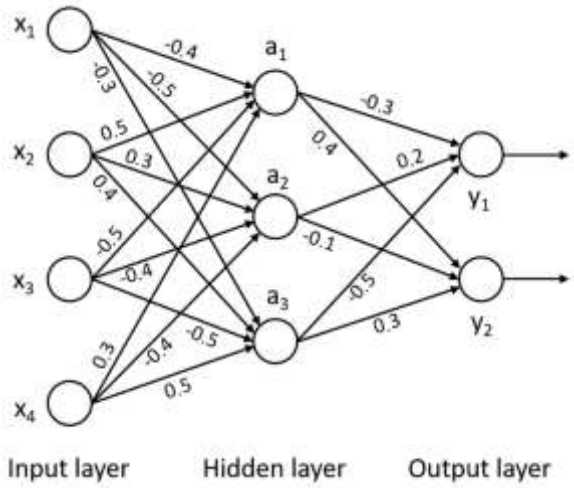

Figure 3. A multi-layer ANN with one hidden-layer

Table 1. Training Data as example

No	y	x1	x2	x3	x4
1	0	0	0	0.5	1
2	0	0.5	0.5	0.5	0.5
3	0	0	0.5	0	1
4	0	0.5	0	1	0
5	0	1	0	1	0.5
6	1	0	0.5	0	1
7	1	0	0.5	0	1
8	1	0	0	1	0.5
9	1	0.5	0.5	0	0.5
10	1	0.5	0.5	1	1

For example, a multi-layer ANN is formed with one hidden layer, as shown in Figure 3. The ANN has four neurons in the input layer, three neurons in the hidden layer, and two neurons in the output layer with initial weights as shown in the figure. Training data is used with five outputs worth 0 and five outputs worth 1 as shown in Table 1. Next, look for the sum of the outputs z_i in the hidden layer, for example, for training data number 1 using the weights in Figure 3 to obtain

$$z_1 = 0 \cdot (-0.4) + 0 \cdot 0.5 + 0.5 \cdot (-0.5) + 1 \cdot 0.3 = 0.05,$$
$$z_2 = 0 \cdot (-0.5) + 0 \cdot 0.3 + 0.5 \cdot (-0.4) + 1 \cdot (-0.4) = 0.2,$$
$$z_3 = 0 \cdot (-0.3) + 0 \cdot 0.4 + 0.5 \cdot (-0.5) + 1 \cdot 0.5 = 0.25.$$

Furthermore, by using the sigmoid function as the activation function, we find

$$a_1 = f(0.05) = \frac{1}{1 + \exp(-0.05)} = 0.512497.$$

In the same way, for training data number 1, $a_2 = 0.549834$ and $a_3 = 0.562177$ are obtained. Then count the number of outputs in the output layer, for example, for training data number 1, it is obtained

$$z_1 = 0.512497 \cdot (-0.3) + 0.549834 \cdot 0.2 + 0.562177 \cdot (-0.5) = -0.32487,$$
$$z_2 = 0.512497 \cdot 0.4 + 0.549834 \cdot (-0.1) + 0.562177 \cdot 0.3 = 0.31867.$$

By using the same activation function, we get $y_1 = 0.419489$ and $y_2 = 0.579$. Then look for the error from the training data output compared to the expected value. In this experiment y_1 is the output $y = 0$ and y_2 is the output $y = 1$. For training data number 1, $d_1 = 1$ and $d_2 = 0$ (because the desired result is $y = 0$), so the error is obtained

$$\varepsilon \triangleq \frac{1}{2}[(1-0.419489)^2+(0-0.579)^2] = 0.336117.$$

The calculation is repeated for training data numbers 2-10; the values of the hidden layer, output layer, and errors are obtained as presented in **Table 2**.

In detail, the Back Propagation algorithm is explained with the following steps:

1. Initialization the first training data.
2. For each training data, compute the weight change to the output layer.

$$\Delta w_{kj}(p) = \eta \Phi_k(p) y_j(p-1) \quad (9)$$

where

$$\Phi_k = y_k(1-y_k)(d_k-y_k) \quad (10)$$

and j represents the *j*-th input to the *k*-neuron at the output layer (p).

3. Do Backpropagation to the *r*-th hidden layer by using formula:

$$\Delta w_{ji}(r) = \eta \Phi_j(r) y_i(r-1) \quad (11)$$

$$\Phi_j(r) = y_j(r)\left[1-y_j(r)\right]\sum_k \Phi_k(r+1) w_{kj}(r+1) \quad (12)$$

where i represents the *i*-th input to *j*-neuron in the *r*-th hidden layer.

4. Repeat Step 2 for $r = p$-1, p-2, ⋯, 2, 1.
5. Calculate the average weight change $\Delta w(m)$ for all training data.
6. Update $w(m+1)$ using $w(m)$ and $\Delta w(m)$ for the $(m+1)$-iteration

$$w_{kj}(m+1) = w_{kj}(m) + \Delta w_{kj}(m) \quad (13)$$

7. Repeat the whole process by applying the next training vector for $(m+2)$, $(m+3)$, ⋯ until the error obtained converges.

Table 2. Value of hidden layer, output layer, and error at the first training

No	a1	a2	a3	y1	y2	error
1	0.512497	0.549834	0.562177	0.419489	0.579	0.336117
2	0.487503	0.475021	0.512497	0.423726	0.574748	0.331213
3	0.634136	0.634136	0.668188	0.401911	0.59645	0.356732
4	0.331812	0.34299	0.34299	0.449562	0.550161	0.30283
5	0.320821	0.331812	0.365864	0.446996	0.551048	0.304734
6	0.634136	0.634136	0.668188	0.401911	0.59645	0.162192
7	0.634136	0.634136	0.668188	0.401911	0.59645	0.162192
8	0.413382	0.450166	0.437823	0.437111	0.562591	0.191197
9	0.549834	0.524979	0.574443	0.414067	0.584134	0.172198
10	0.462570	0.475021	0.512497	0.425554	0.572308	0.182008

Using the output in **Table 2**, look for changes in weight to the output layer. For example, using data training number 1, for neuron y_1 in the output layer

$$\Phi_1 = (0.419489)(1-0.419489)(1-0.419489) = 0.141365$$

until obtained $\Delta w_{11}(3) = 0.144898$, $\Delta w_{12}(3) = 0.155454$, and $\Delta w_{13}(3) = 0.158944$. Next, back propagation is done to the hidden layer. For example, using training data number 1, for neuron a1 in the hidden layer

$$\Phi_1(2) = (0.512497)[1-0.512497][(0.141365)(0.144898)+(-0.141136)(-0.144664)] = 0.010219$$

until obtained $\Delta w_{11}(2) = 0$, $\Delta w_{12}(2) = 0$, $\Delta w_{13}(2) = 0.005109$, and $\Delta w_{14}(2) = 0.010219$. Because in this example there is only 1 hidden layer, the back propagation algorithm for the first iteration produces a new weight as stated in **Table 3**. After repeating the back propagation algorithm for 1010 iterations, the average error converges to 0.21247 and the final weight is expressed as in **Table 4**.

Table 3. New weight after first iteration

w_{ji}	a1	a2	a3
x1	-0.3982	-0.49817	-0.29933
x2	0.502168	0.302152	0.401257
x3	-0.49698	-0.39683	-0.49859
x4	0.305092	0.405188	0.503415

w_{kj}	y1	y2
a1	-0.28982	0.390178
a2	0.210943	-0.11058
a3	-0.48857	0.288955

Table 4. Final weight after 1010 iterations

w_{ji}	a1	a2	a3
x1	0.757571	0.673071	-0.60242
x2	1.615528	1.433603	2.655797
x3	1.528093	1.590279	0.045318
x4	2.987346	3.061566	4.464579

w_{kj}	y1	y2
a1	1.685363	-1.65204
a2	2.451821	-2.42621
a3	-4.38744	4.326629

In use, of course, the neural network method is not only used to classify 1 datum but a group of datums known as test data. To measure the performance of the classification algorithm used confusion matrix. In the case of classification of 2 classes, in the confusion matrix, there is information that can be used to compare the results of the classification carried out by the ANN/LR method with the actual classification. In this case, there are 4 terms used, namely True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Each can be interpreted as follows: TP is the number of positive datums that are correctly classified by the ANN/LR method, TN is the number of negative datums that are correctly classified by the ANN/LR method, FN is the number of negative datums but classified incorrectly by the ANN/LR method and FP is the number of datums positive but classified incorrectly by the ANN/LR method. In tabular form, this can be stated in **Table 5**. Furthermore, the accuracy of the ANN/LR method is formulated as:

$$A = \frac{TP + TN}{TP + TN + FP + FN},$$

which is the ratio between the correct prediction and the overall data, precision is formulated as:

$$P = \frac{TP}{TP + FP},$$

that is, the ratio between a positive correct prediction and the overall positive predicted outcome. Sensitivity (recall) is formulated as:

$$R = \frac{TP}{TP + FN},$$

which is the ratio between positive correct predictions and positive overall data and the last F1 score is formulated as:

$$F1 = 2\frac{P * R}{P + R}.$$

which is a comparison of the average precision and recall [14]. Accuracy has almost the same value as F1. If the classification is more than two classes, the accuracy can be obtained from the comparison between the number of diagonal elements in the confusion matrix divided by the total number of elements in the confusion matrix, but the F1 value cannot be determined in this case.

Table 5. Table of Classification

Class	Classified Positive	Classified Negative
Positive	TP	FN
Negative	FP	TN

The data used in this study is asthma data contained in the R program package, namely SNPassoc [17]. Asthma data has 1578 rows and 57 columns obtained from 1578 individuals and epidemiological variables, namely country, gender, age, BMI, smoke status, and case/control, as well as 51 SNP from 1578 individuals. The data will be classified as case/control status of asthma based on SNP, which is closely related to the disease. In this case, there were 340 cases and 1238 controls. The data were obtained from Australia, Belgium, Estonia, France, Germany, Norway, Spain, Sweden, Switzerland and the UK (United Kingdom), respectively 127, 14, 6, 219, 154, 177, 377, 281, 100, and 123. It will be compared between the logistic regression method and the neural network method, which provides higher accuracy, precision, recall, and F1 score.

3. RESULTS AND DISCUSSION

Based on the asthma data, individuals who did not have missing data were selected so that 1091 individuals who were free of missing data would be obtained. SNP genetic data was transformed using codominant genetic traits, namely for genetic data AA, AC, CC were given a score of 0, 0.5, and 1, respectively, while CC, CT and TT were scored 0, 0.5, and 1, respectively. The scoring is based on the smallest alphabetical order given a low score. Likewise, for the other genetic codes, scores are given analogously. In this case, there were 235 cases and 856 controls. Furthermore, by using these data, the logistic regression method is used to predict the class of each individual, which is included in the special or control. Table 6 states the average results of accuracy, precision, recall, and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30%, and 40%. In this case, one hidden layer is used on the ANN and repeated $B = 1000$ times because the selection of which individuals are part of the training data and which individuals are part of the test data can be done arbitrarily. It can be seen that the average accuracy, precision, and F1 score in the logistic regression method is higher than the ANN method. However, the recall value of the ANN method gives better results. The difference is significant when tested using the Kolmogorov-Smirnov statistic. Likewise, the difference in average accuracy, precision, recall, and F1 scores for the use of variations in the proportion of test data 10%, 20%, 30%, and 40% gives results that are not too far away but differ significantly when tested using the Kolmogorov-Smirnov statistic.

Table 6. The results of the average accuracy, precision, recall, and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30% and 40% using the logistic regression (LR) method.

Proportion of Testing Data	Accuracy	Precision	Recall	F1 score
10 %	0.7726	0.9820	0.7829	0.8707
20 %	0.7711	0.9765	0.7844	0.8697
30 %	0.7680	0.9707	0.7845	0.8675
40 %	0.7637	0.9615	0.7855	0.8645

Table 7. The results of the average accuracy, precision, recall, and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30%, and 40% using the ANN method.

Proportion of Testing Data	Accuracy	Precision	Recall	F1 score
10 %	0.7652	0.7851	0.9650	0.8649
20 %	0.7635	0.7851	0.9622	0.8639
30 %	0.7607	0.7848	0.9577	0.8619
40 %	0.7686	0.7848	0.9525	0.8595

Figure 4 and Figure 5, respectively, present histograms of accuracy, precision, recall, and F1 score if 20% of the test data are used and when the logistic regression method and the ANN method are used. It can be seen in Figure 4 that the histogram of the values of accuracy, precision, and F1 score tends to skew to the

left, while the recall histogram tends to be symmetrical. This is also supported by the p-values of the normality test of the values, namely 0.0344, 0.0000, 0.5047, and 0.27932 so that the histogram values are not normally distributed except for histogram recall and F1 score. Likewise, in **Figure 5**, it can be seen that the histograms of accuracy, recall, and F1 scores tend to be skewed to the left, while the precision histograms tend to be symmetrical. This is also supported by the *p*-values of the normality test of the values, namely 0.0000, 0.684, 0.0000, and 0.0000 so that the histogram values are not normally distributed except for the histogram of precision values. **Figure 6** shows a boxplot comparison between the values of accuracy, precision, recall, and F1 score when the logistic regression method and the ANN method are used. It can be seen that the median accuracy and F1 score differ only relatively small but significantly different.

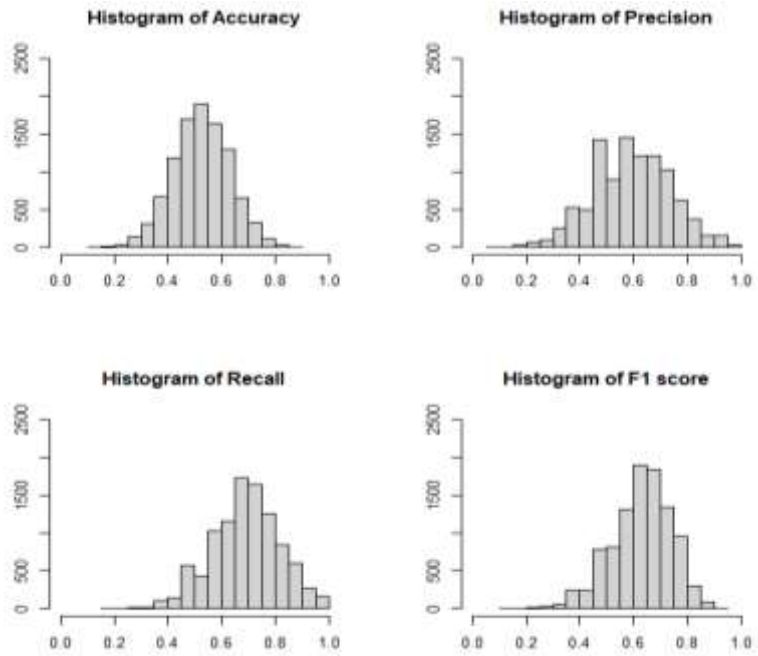

Figure 4. Histogram of Accuracy, Precision, Recall and F1 score if 20% test data and logistic regression method are used.

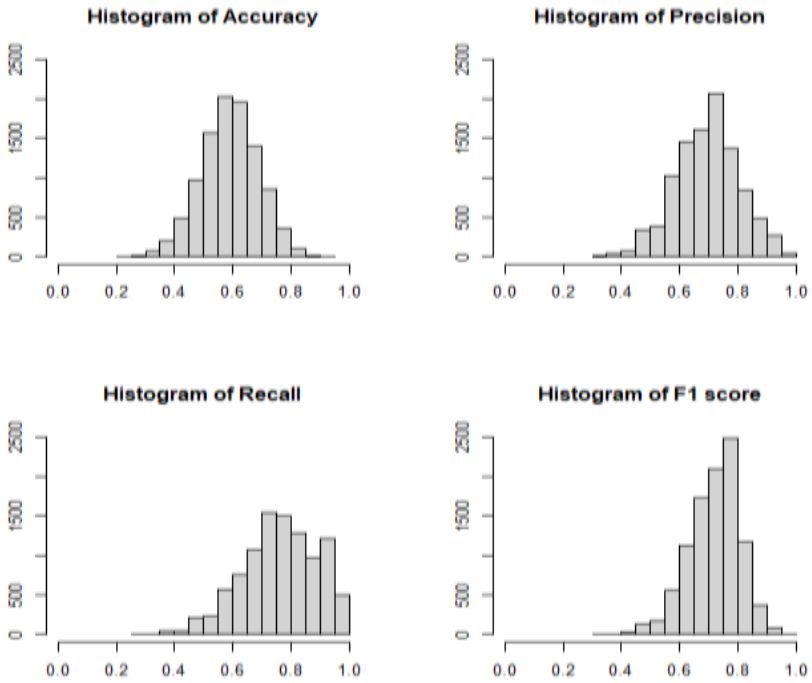

Figure 5. Histogram of Accuracy, Precision, Recall, and F1 score if 20% test data and ANN method are used.

Based on the asthma data, individuals who did not have missing data were selected and those who had a positive correlation with the case and control classes were selected so that 1091 individuals and 22 SNPs

were obtained. Furthermore, by using these data, the logistic regression method was used to predict the class of each individual that was included in the case or control. Table 7 states the average results of accuracy, precision, recall, and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30%, and 40%. In this case, one hidden layer is used in the ANN.

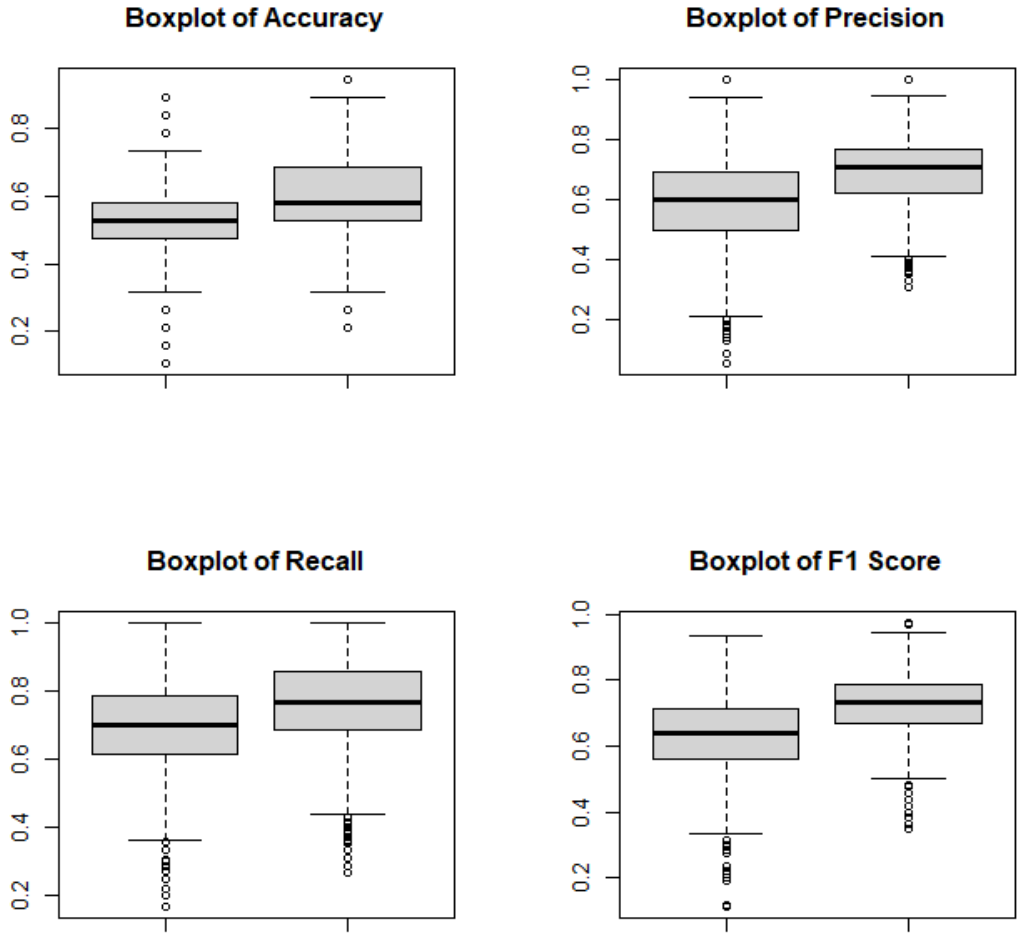

Figure 6. Boxplot Comparison of the values of accuracy, precision, recall, and F1 score if 20% of test data is used with logistic regression (left) and ANN (right) methods.

Table 8. The results of the average accuracy, precision, recall, and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30%, and 40% using the logistic regression method.

Proportion of Testing Data	Accuracy	Precision	Recall	F1 score
10 %	0.7830	0.9947	0.7860	0.8776
20 %	0.7805	0.9938	0.7840	0.8769
30 %	0.7803	0.9930	0.7843	0.8762
40 %	0.7808	0.9914	0.7856	0.8765

Table 9. The results of the average accuracy, precision, recall and F1 score if the genetic code is used with variations in the proportion of test data 10%, 20%, 30% and 40% using the ANN method.

Proportion of Testing Data	Accuracy	Precision	Recall	F1 score
10 %	0.7747	0.7825	0.9862	0.8720
20 %	0.7756	0.7844	0.9844	0.8728
30 %	0.7760	0.7860	0.9826	0.8731
40 %	0.7730	0.7845	0.9799	0.8711

Table 10. The Results of the average accuracy, precision, recall and F1 score if the genetic code is used with a variation of the proportion of test data 20%, using the logistic regression method and the ANN method (hidden neuron = 1).

Country	Accuracy (LR)	Accuracy (ANN)	Precision (LR)	Precision (ANN)	Recall (LR)	Recall (ANN)	F1 score (LR)	F1 score (ANN)
Australia	0.5210	0.7672	0.5700	0.8043	0.7805	0.9541	0.6555	0.8535
France	0.6951	0.8454	0.7644	0.8849	0.8781	0.9502	0.8138	0.9138
Germany	0.8088	0.9552	0.8416	0.9617	0.9574	0.9932	0.8923	0.9767
Norway	0.7475	0.7804	0.9539	0.9599	0.9536	0.9938	0.8502	0.9757
Spain	0.7708	0.8220	0.8622	0.8700	0.8719	0.9357	0.8656	0.8997
Sweden	0.5933	0.5509	0.7196	0.6466	0.6799	0.6944	0.6946	0.6479
Switzerland	0.5253	0.6063	0.5973	0.7049	0.6974	0.7790	0.6327	0.7258
UK	0.4545	0.4876	0.4937	0.5276	0.5197	0.5591	0.4942	0.5196

Table 10, Table 11 and Table 12, respectively, present the results of the average accuracy, precision, recall, and F1 score when 20% test data is used using the logistic regression method and the ANN method when hidden neuron = 1, 5, and 10 are used. In this study, the individuals used were from Australia, Belgium, Estonia, France, Germany, Norway, Spain, Sweden, Switzerland, and there were 20, 152, 132, 74, 279, 217, 91 and 107 individuals, respectively. Furthermore, a separate analysis was also performed with B = 10000 replicates. By using the Kolmogorov-Smirnov test, it can be found that there is a significant difference between accuracy, precision, recall and F_1 score obtained by the logistic regression method compared to the ANN method. Furthermore, it was found that the accuracy, precision, recall and F_1 in the ANN method were higher than using the linear regression method.

Table 11. The results of the average accuracy, precision, recall and F1 score if the genetic code is used with a variation of the proportion of test data 20%, using the logistic regression method and the ANN method (hidden neuron = 5).

Country	Accuracy (LR)	Accuracy (ANN)	Precision (LR)	Precision (ANN)	Recall (LR)	Recall (ANN)	F1 score (LR)	F1 score (ANN)
Australia	0.5246	0.7478	0.5736	0.8011	0.7821	0.9310	0.6582	0.8402
France	0.6961	0.8009	0.7653	0.8827	0.8784	0.8962	0.8145	0.8872
Germany	0.8107	0.9390	0.8431	0.9616	0.9580	0.9762	0.8935	0.9681
Norway	0.7468	0.9440	0.7798	0.9589	0.9534	0.9846	0.8497	0.9705
Spain	0.7712	0.7843	0.8619	0.8704	0.8729	0.8841	0.8659	0.8757
Sweden	0.5938	0.5463	0.7198	0.6547	0.6805	0.6558	0.6952	0.6500
Switzerland	0.5266	0.5769	0.5992	0.6997	0.6976	0.7268	0.6340	0.7029
UK	0.4737	0.4824	0.4955	0.5264	0.5188	0.5278	0.4946	0.5141

Table 12. The results of the average accuracy, precision, recall and F1 score if the genetic code is used with a variation of the proportion of test data 20%, using the logistic regression method and the ANN method (hidden neuron = 10).

Country	Accuracy (LR)	Accuracy (ANN)	Precision (LR)	Precision (ANN)	Recall (LR)	Recall (ANN)	F1 score (LR)	F1 score (ANN)
Australia	0.5226	0.7474	0.5700	0.7974	0.7777	0.9335	0.6584	0.8398
France	0.6951	0.8191	0.7651	0.8851	0.8777	0.9169	0.8140	0.8989
Germany	0.8109	0.9450	0.8440	0.9616	0.9573	0.9825	0.8937	0.9713
Norway	0.7455	0.9487	0.7789	0.9590	0.9532	0.9893	0.8490	0.9731
Spain	0.7721	0.7942	0.8627	0.8693	0.8732	0.8987	0.8666	0.8825
Sweden	0.5924	0.5519	0.7178	0.6564	0.6797	0.6694	0.6937	0.6580
Switzerland	0.5266	0.5845	0.5993	0.7000	0.6984	0.7454	0.6344	0.7124
UK	0.4742	0.4780	0.4941	0.5232	0.5202	0.5263	0.4948	0.5113

Research related to this research is in research [18] on the use of machine learning, namely the integration of RF-SVM, which produces accuracy, precision, and recall, respectively 62.5%, 65.3%, and 69%. This result is slightly lower than the result obtained when both logistic regression and ANN methods are used. Other studies on the use of neural networks are also included in the study [19], but the results obtained are not compared with other methods. The accuracy of this study even reached 96.23%, but there is no information about other classification goodness measures such as precision, recall, and F1 score. Other research on the use of machine learning, namely the SVM, Naïve Bayes, and Decision Tree methods can be found in the paper [20]. In this paper, the accuracy results are 69%, 67%, and 68% respectively for the SVM, Naive Bayes, and Decision Tree methods. The use of ANN in determining accuracy, sensitivity, and specifications is contained in the paper [21]. Obtained respectively 67.5%, 62.16%, and 70.73% for accuracy, sensitivity and specification. Furthermore, another study on the hybrid between SVM and ANN methods was carried out in paper [22] and gave an accuracy of 98.08% using 25% test data. This paper also explains in detail about accuracy, precision, recall, and F1 score, but no other variation of test data is carried out.

4. CONCLUSIONS

In this paper, it has explained how to use the ANN method on asthma data in the R program package, namely ***SNPassoc***. The results obtained were compared with the logistic regression method. The following results were obtained:

1. When all information from various countries is used, the logistic regression method gives higher average accuracy, precision, and F1 scores than the ANN method, but the average recall applies the other way around.
2. When a separate analysis is performed for each country, the logistic regression method gives higher accuracy, precision, recall, and F1 scores in the ANN method compared to the logistic regression method.

In this study, BMI (Body Mass Index) predictions can be developed using the same data as the ANN regression method and similarly, deep learning methods can be used to analyze the same data.

REFERENCES

[1] H. Sartor *et al.*, "The Association of Single Nucleotide Polymorphisms (SNPs) with Breast Density and Breast Cancer Survival: the Malmö Diet and Cancer Study," *Acta Radiologica*, vol. 61, no. 10, pp. 1326–1334, 2020.
[2] M. N. Mikhail, A. Y. Sayed, M. S. Mabrouk, and A. M. Eldeib, "Investigation of Genome-Wide Association SNPs and Alzheimer ' s Disease," *American Journal of Biomedical Engineering*, vol. 10, no. 1, pp. 1–8, 2020.
[3] D. saber Morgan, R. A. Mohamed, M. M. Abdelkhalek, and A. A. Mohamed, "Detection of Single Nucleotide Polymorphism (SNP) (rs34819629) and its Association with Pediatric Type 1 Diabetes Mellitus Dalia," *Egyptian Journal of Medical Research (EJMR)*, vol. 3, no. 2, pp. 185–195, 2022.
[4] Y. Tursinawati, R. F. Hakim, A. Rohmani, A. Kartikadewi, and F. Sandra, "CAPN10 SNP-19 is Asociated with Susceptibility of Type 2 Diabetes Mellitus: A Javanese Case-Control Study," *Indonesian Biomedical Journal*, vol. 12, no. 2, pp. 109–114, 2020.
[5] Y. C. Kim *et al.*, "Genome-Wide Association Study Identifies Eight Novel Loci for Susceptibility of Scrub Typhus and Highlights Immune-Related Signaling Pathways in Its Pathogenesis," *Cells*, vol. 10, no. 3, 2021.
[6] C. A. C. Montañez, P. Fergus, A. C. Montañez, A. Hussain, D. Al-Jumeily, and C. Chalmers, "Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs," 2018.
[7] Y. Tomita *et al.*, "Artificial Neural Network Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma," *BMC Bioinformatics*, vol. 5, no. 1, 2004.
[8] J. Stangierski, D. Weiss, and A. Kaczmarek, "Multiple Regression Models and Artificial Neural Network (ANN) as Prediction Tools of Changes in Overall Quality during the Storage of Spreadable Processed Gouda Cheese," *European Food Research and Technology*, vol. 245, no. 11, pp. 2539–2547, 2019.
[9] V. Gahlaut, V. Jaiswal, S. Singh, H. S. Balyan, and P. K. Gupta, "Multi-Locus Genome Wide Association Mapping for Yield and Its Contributing Traits in Hexaploid Wheat under Different Water Regimes," *Scientific Reports*, vol. 9, no. 1, 2019.
[10] P. Schober and T. R. Vetter, "Linear Regression in Medical Research," *Anesthesia & Analgesia*, vol. 132, no. 1, pp. 108–109, 2019.
[11] S. Ghosal, S. Sengupta, M. Majumder, and B. Sinha, "Linear Regression Analysis to Predict the Number of Deaths in India due to SARS-CoV-2 at 6 Weeks from Day 0 (100 Cases - March 14th 2020)," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 4, pp. 311–315, 2020.
[12] S. I. Bangdiwala, "Regression: Binary Logistic," *International Journal of Injury Control and Safety Promotion*, vol. 25, no. 3, pp. 336–338, 2018.
[13] N. Srimaneekarn, A. Hayter, W. Liu, and C. Tantipoj, "Binary Response Analysis using Logistic Regression in Dentistry,"

International Journal of Dentistry, vol. 2022, 2022.
[14] D. Graupe, *Principles of Artificial Neural Networks*, 3rd ed. Jurong East: World Scientific Publishing Co. Pte. Ltd. All, 2013.
[15] M. L. Minsky and S. A. Papert, *Perceptrons, Reissue of the 1988 Expanded Edition with a new foreword by Léon Bottou: An Introduction to Computational Geometry*. Cambridge: MIT Press, 2017.
[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," San Diego La, 1986.
[17] J. R. González *et al.*, "SNPassoc: An R Package to Perform Whole Genome Association Studies," *Bioinformatics*, vol. 23, no. 5, pp. 644–645, 2007.
[18] J. Gaudillo *et al.*, "Machine Learning Approach to Single Nucleotide Polymorphism-based Asthma Prediction," *PLOS ONE Journal*, vol. 14, no. 12, 2019.
[19] H. Soumare, S. Rezgui, N. Gmati, and A. Benkahla, "New Neural Network Classification Method for Individuals Ancestry Prediction from SNPs data," *BioData Mining*, vol. 14, no. 1, 2021.
[20] F. J. Shaikh and D. S. Rao, "Prediction of Cancer Disease using Machine Learning Approach," *Materials Today: Proceedings*, vol. 50, pp. 40–47, 2022.
[21] L. Besic, I. Muhovic, A. Asic, A. Catic, L. Gurbeta, and A. Badnjevic, "Application of Neural Networks to the Prediction of a Phenotypic Trait of Pacific Lampreys based on Single Nucleotide Polymorphism (SNP) Genetic Markers," *Biomedical Research and Clinical Practice*, vol. 2, no. 5, pp. 1–7, 2017.
[22] P. Nanglia, S. Kumar, A. N. Mahajan, P. Singh, and D. Rathee, "A Hybrid Algorithm for Lung Cancer Classification using SVM and Neural Networks," *ICT Express*, vol. 7, no. 3, pp. 335–341, 2021.