Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

A Backpropagation Artificial Neural Network Approach for Loan Status Prediction

Edwin Setiawan Nugraha[1], Gabrielle Jovanie Sitepu [2]

[1,2]Actuarial Science Study Program, Faculty of Business, President University
Jl. Ki Hajar Dewantara, Mekarmukti, Cikarang Utara, Bekasi Regency, West Java, 17550, Indonesia
E-mail: [1]edwin.nugraha@president.ac.id, [2]gabrielle.sitepu@student.president.ac.id

Article:
Accepted: July 22, 2022
Revised: May 02, 2022
Issued: November 15, 2022

***Correspondence Address:**
edwin.nugraha@president.ac.id

ABSTRACT

Providing credit has become the main source of profit for financial and non-financial institutions. However, this transaction might lead to credit risk where debtors are unable to complete their obligations. In this case, the prediction of loan status is extremely important to minimize the risks. The objective of this work is to predict loan status by using a backpropagation algorithm. The used dataset consists of 1 dependent variable and 13 independent variables which 75 % are for data training and 25 % for data testing. There are two main simulation experiments namely simulation involving all predictor variables and another one involving just only predictor variables has a significant relationship with the target variable. The first main simulation experiment shows the best performance metrics from the first model are 94.37% accuracy, 78.57% sensitivity, 98.25% specificity, 91.67% precision, and 84.62% F1 score. The performance metrics of the second one are the same as the best performance metrics of the first simulation. The results of this study can potentially be applied by financial institutions to assist in the feasibility assessment of prospective debtors to reduce company losses.

Keywords: *Credit Risk, Loan Status, Backpropagation, Artificial Neural Network*

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

I. INTRODUCTION

Every business is aiming for profit, financial institutions such as banks included. Bank as a company collects money from the public with savings and redistributes money to a certain community with credit [1]. Providing credit has become the main activity of banks since this transaction generates large profits for the institution. Credit is an arrangement in which a party or other instance enables its client to future repayment for its money, goods, property, or services supplied or borrowed [2]. However, this lending and borrowing transaction in financial institutions has a chance to turn into a huge source of loss. The issues arise when creditors' position is vulnerable since the amount of transaction of credits occurred but the creditors aren't cautious in examining the future debtor. Banks are exposed to the credit risk of failure to observe the debtors' capability to complete their obligations.

Credit risk is referred to as the probability of default in a loan agreement. The risk occurred of the increasing possibility of irrecoverable loans because of the case of outright default [3]. An irrecoverable loan is an incident when a debtor is unable to repay its loan installment within a given period. The arising danger of credit risk declines the value of company assets and in an extreme way might lead to insolvency of the bank [3]. It is the obvious bank would reject the idea of company bankruptcy because of the debtor's irresponsibility. In order to prevent this incident, the bank is required to obligate a noble credit risk management. There are steps implemented by the bank before accepting a credit application, one of them is applicant assessment. The applicant's assessment allows the bank to determine the eligibility and ability of prospective credit recipients to repay their obligations. Thus, the aim is to minimize the credit risk by assessing the loan status of the prospective customer through the evaluation process in order to escape unexpected events that might inflict financial loss.

The problem to solve with the applicant's assessment is to distinguish potential customers within the categories that are eligible for a loan and not. Eligible customers will have their loan application addressed, while the bank has the right to deny those customers non-eligible. Another word for an applicant's assessment is to predict the status of a loan application between accepted or rejected. In the application, there has been a traditional statistical method developed to obtain an accurate and guaranteed prediction result. Some examples of developed methods are Logistic Regression and Linear Discriminant Analysis. Several studies applied traditional methods in predicting the loan status including: [4] building a prediction model to classify the loan status into accepted or rejected. This study conducted by using Logistic Regression and Naïve Bayes classifier, with a result of model performance is scaled by accuracy obtained at 85.9% and 84.62% respectively. A study conducted by [5] utilizing Logistic Regression in classifying loan status into accepted or rejected has received an accuracy of 81%. Another study by [6] in building a prediction model to predict loan safety apply Logistic Regression has obtained the best result with 81.11% accuracy. Previous studies show that model performance is good but still less than 90%.

Furthermore, with the advancement of technology, an information processing system is presented for solving the problem with help of a machine. Machine learning is introduced for solving a typical classification problem. Some studies implied that machine learning has great capability in classifying and can replace the use of traditional statistical methods [7], [8]. Machine learning is widely used for loan status prediction, for example [9] builds a loan default prediction model using Random Forest algorithm with accuracy of 98% and Decision Tree algorithm with accuracy of 95%.

Artificial Neural Network (ANN) is one of machine learning methods widely used in today's age. Research of predicting and classifying has been conducted and the result found that Neural Networks if compared with Logistic Regression, will show higher accuracy and ROC values, therefore, Neural Network outperform Logistic Regression [7], [10], [11]. Furthermore, this research applies one of algorithms in ANN namely Backpropagation. Backpropagation is an algorithm known for its ability to obtain minimal error and generate output that is closer to desired output with every forward pass [12]. This ability made Backpropagation a great proposed model for classification problems. Another advantage of Backpropagation is simple and easy to construct

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

the program and works well with such complex datasets [13],[14]. Besides of number of inputs, there is no complex parameter in Backpropagation that must be calculated and the application of Backpropagation doesn't require prior knowledge of the network making it convenient for use [14]. These advantages made Backpropagation a fast-learning convergence for classification. Despite these advantages, there is a limited amount of research applying Backpropagation for loan status prediction, this paper is expected to contribute to delivering knowledge to the audience.

This study aims to construct an applicant assessment to predict the accepted or rejected loan. The method proposed for predicting the loan status is the theory of Backpropagation. This study builds a prediction model with Backpropagation utilizing available historical data and determining which loan should be accepted or rejected. Later, the prediction result is used to conclude the performance of Backpropagation in classifying applicant loans.

The structure of this paper is as follows. Section II provides an explanation of Backpropagation and provides a list of information on the method of analyzing the data. Section III demonstrate the process of generating a model for this research and displays the research result obtained. Section IV is the conclusion of this study.

II. METHODOLOGY

This study applies Backpropagation to build a classification model for loan status. Backpropagation is known for its competency in recognizing data patterns and minimizing output error by optimizing the value of model parameters.

First step for building the prediction model with Backpropagation is to collect the dataset. Obtained dataset processed into transformation aims to improve the quality of dataset. This transformation process is called data pre-processing, total of six steps will further explain. One noticeable step in pre-processing is feature selection, where for this study two data models are formed with purpose to prove models' performance in prediction considering data with fewer or more variables. Immediately after preprocessing, data will use for training model with Backpropagation method. Activation function is sigmoid function that is suitable for binary classification as the research purpose.

2.1 Data Collection

This research uses secondary data obtained from online database github.com [15] which has 983 observations. The number of samples for this research is taken with Yamane's formula. There is total of 12 variables within dataset. From those variables, there are eleven (11) independent variables that act as the predictor and only one (1) target variable, which is variable "Loan Status". List of variables in dataset is summarized in Table 1 below.

Table 1. Data variable

	Variable	Description
X_1	Gender	Categorical
X_2	Married	Categorical
X_3	Dependents	Categorical
X_4	Education	Categorical
X_5	Self Employed	Categorical
X_6	Applicant Income	Numerical
X_7	Co-applicant Income	Numerical
X_8	Loan Amount	Numerical
X_9	Credit History	Categorical
X_{10}	Loan Term	Categorical
X_{11}	Property Area	Categorical
Y_1	Loan Status	Categorical

The descriptive statistic of data variables for categorical variables is shown in Table 2, and for numerical variables is shown in Table 3.

2.2 Data Preprocessing

Data Preprocessing is conducted to enhance the quality of data and improve model performance [16]. There are six data preprocessing in this study.

Table 2. Descriptive statistics of categorical variables

Variables	Category	Frequency
Gender	Male	624
	Female	145
Married	Yes	498
	No	271
Dependent	0	441
	1	122
	2	135
	3+	71
Education	Graduate	607
	Not Graduate	162

Self Employed	Yes	98
	No	671
Loan Term	36	3
	60	2
	84	5
	120	4
	180	55
	240	4
	300	15
	360	659
	480	19
Credit History	1	653
	0	116
Property Area	Rural	228
	Urban	263
	Semiurban	278
Loan Status	Y	561
	N	208

Table 3. Descriptive statistics of numerical variables

	Mean	Median	Mode
Applicant Income	5091.061	3850	2500
Co-applicant Income	1561.239	1032	0
Loan Amount	141.750	128	110

2.2.1 Data Cleaning

The missing data expressed with null data or "NA" is eliminated by deleting the missing information [16]. After the elimination, the original dataset of 983 data turns into 769 data.

2.2.2 Feature Selection

Feature selection reduces the number of input variables by removing non – informative predictors and maintains the most informative predictor variables [17]. This research can obtain two data models, where one consists of variable within dataset means more input and second consists of informative variables means less input. This study checks Backpropagation capability of predicting with more input or less input. Since predictor in dataset consists of numerical and categorical variables while target variable is a binary variable, supervised method was applied to remove the irrelevant variables based on their relationship with target variable [17]. Point Biserial Correlation Coefficient is used to check correlation between numerical variables and target variables. To measure the Point Biserial Correlation Coefficient value such formula is seen in Equation (1)[18]:

$$r_{pbis} = \frac{M_p - M_T}{SD_T}\sqrt{\frac{p}{q}} \quad (1)$$

where M_T is mean score total of numerical variables, M_p is mean score from population has possibility of an accepted loan, SD_T is standard deviation total of numerical variables, p is probability of loan status accepted within the population, and q is probability of loan status rejected. Chi–square Test was used to check correlation between categorical variables. Chi-square value obtained from formula in Equation (2)[19]:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

where O is the observed data and E is the expected data.

2.2.3 Data Encoding

Machine learning required the input and output variables to be numeric which means every categorical data should be encoded into the numerical label. Research applied label encoding for binary variables and dummy encoding for non-binary categorical variables.

2.2.4 Data Splitting

The sample size for this study is generated by computing Yamane's sample in Equation (3)[20]:

$$n = \frac{N}{1 + N(e)^2} \quad (3)$$

From Yamane's, the sample size obtain is 284 data. After setting sample size, divide data into training and testing datasets. The splitting ratio chosen in this study is 75% for training dataset and 25% for testing dataset. There is total of 213 data for training and 71 data for testing.

2.2.5 Data Normalization

Data normalization is applied for numerical variables [16]. Previous step of data encoding transforms categorical variables into numerical values of 0 and 1. Numerical variables have a larger range between their values with the remaining categorical variable values. Since the dataset has different ranges, normalization is needed to scale data as fall within smaller range such as -1 to 1 or 0 to 1 [16]. The normalization conducted with Min and Max Normalization is transforming the data where minimum value becomes 0, maximum value becomes 1, and the other value becomes

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

decimal number between 0 and 1. Apply Min and Max Normalization, with Equation (4)[16]:

$$x' = \frac{x - x_{max}}{x_{max} - x_{min}} \quad (4)$$

where x_{max} maximum value of variable, x_{min} is minimum value of variable and x is the value of variables.

2.2.6 Outlier Filtering

Noisy data might interrupt information processing. Filtering the outlier with Interquartile Ratio (IQR) and the values of IQR will determine the change in the outlier value and create zero outliers within dataset.

2.3 Backpropagation Algorithm

This section explains how to solve a problem with Backpropagation. The training process required three phases: Feed-forward propagation, Backpropagation, and Weight Adjustment [21]. These phases are defined as:

a. Feedforward: Feedforward is the process of inserting the inputs into the network and obtaining the output of model.
b. Backward Propagation: This process of finding errors of weights and biases carries out in backward direction, therefore it starts from output layer (error of the output) to hidden layer (error of weight and bias from hidden units to output unit) and lastly input layer (error of weight and bias from input units to output unit).
c. Weight Adjustment: Updating value of weight and bias from error adjustment of both weight and bias. The new value of weight and bias later are used for the final model.

For the testing process, feed-forward is the only phase since the appropriate parameter has been obtained in training process.

In detail, the training process of Backpropagation is explained further step by step [21]:

Step 0: Initialize weight with small random number.
Determine the "STOP" condition by calculating the target error and maximum epoch.
Step 1: If "STOP" condition at faults, extend to Step 2 – 9.
Step 2: Continue next step for every data

Phase I: Feed-forward Propagation

Step 3: Every input unit $X_i, (i = 1, \dots, n)$ receive input signal $x_i$will deliver those signals toward the units in the next layer (hidden layer).

Step 4: Every hidden unit $Z_j, (j = 1, \dots, p)$ calculated each weight for each input signal (including the bias).

$$z_{in_j} = v_{0j} + \sum_{i=1}^{n} x_i v_{ij} \quad (5)$$

Later, to calculate the output signal in hidden layers used an activation function that has been determined before:

$$z_j = f(z_{in_j}) \quad (6)$$

Moving forward, hidden layer will send these signals toward unit in output layer.

Step 5: For every output unit $Y_k, k = 1, \dots, m$ will calculate each weight for each input signal (including the bias).

$$y_{in_k} = w_{0k} + \sum_{j=1}^{p} z_j w_{jk} \quad (7)$$

Later, to calculate the output signal in output layers used an activation function that has been determined before:

$$y_k = f(y_{in_k}) \quad (8)$$

Moving forward, output layer will send these signals toward unit in output layer.

Phase II: Backward Propagation

Step 6: Every output unit $Y_k, k = 1, \dots, m$ will receive a target pattern that matches the training input pattern and calculates the error between the target and the output generated by the network.

$$\delta_k = (t_k - y_k) f'(y_{in_k}) \quad (9)$$

Factor δ_k is going to be used to calculate the error adjustment Δw_{jk} that later will used to updatew_{jk}, where:

$$\Delta w_{jk} = \alpha \delta_k z_j \quad (10)$$

Later, the bias adjustment Δw_{0k} is calculated and used to correct the value of $w_{0k}$where:

$$\Delta w_{0k} = \alpha \delta_k \quad (11)$$

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

After that, factor δ_k is going to be sent toward under layer (or the layer below) which is hidden layer.

Step 7: For every output unit $Z_j, (j = 1, \dots, p)$ will delta input previously worked in Step 6.

$$\delta_{in_j} = \sum_{k=1}^{m} \delta_k w_{jk} \quad (12)$$

Then multiplied this delta input with activation function to calculate the error information.

$$\delta_j = \delta\_in_j f'(z_{in_j}) \quad (13)$$

Factor δ_j is going to be used to calculate the error adjustment Δv_{ij} that later will used to update v_{ij}, where:

$$\Delta v_{ij} = \alpha \delta_j x_i \quad (14)$$

Later, the bias adjustment Δv_{0j} is calculated and used to correct the value of v_{0j}, where:

$$\Delta v_{0j} = \alpha \delta_j \quad (15)$$

Phase III: Weight Adjustment

Step 8: Every output unit $Y_k, k = 1, \dots, m$ adjust their weights and bias from every hidden unit $j = 0, .., p$, where:

$$w'_{jk} = w_{jk} + \Delta w_{jk} \quad (16)$$

Same way with each hidden output $Z_j, (j = 1, \dots, p)$ will also adjust and update their weights and bias from each input unit $= 0, \dots, n$, where:

$$v'_{ij} = v_{ij} + \Delta v_{ij} \quad (17)$$

Step 9: Test the "STOP" condition

Symbol in equation defined as:

- t is the target output
- X_i is the input unit i
- v_{0j} is the bias on hidden j
- Z_j is the hidden unit j
- w_{0k} is the bias on output k
- Y_k is the output unit k
- α is learning Rate
- δ_k is the portion of error correction weight adjustment for w_{jk}
- δ_j is the portion of error correction weight adjustment for v_{ij}

2.4 Activation Function

Activation function in this research is Sigmoid Function. Sigmoid function has range of $x = (0,1)$ to return the probability and is very suitable for solving classification problems [22]. Equation (17)[22] is the formula of sigmoid:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (18)$$

The derivative of this function is implemented in Backward Propagation phase to find the error adjustment of weight and bias, shown in Equation (18)[22]:

$$f'(x) = Sigmoid \times (1 - Sigmoid) \quad (19)$$

2.5 Loss Function

Calculate the loss of model training for binary classification problem with Binary Cross Entropy. Binary Cross Entropy has function defined in Equation(19)[23]:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) - (1 - y_i)\log(1 - p_i)) \quad (20)$$

2.6 Confusion Matrix

Confusion Matrix is a performance evaluation for machine learning classification for binary classification or multi-class classification [4]. Confusion matrix has four classification terms, TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). These four terms are shown in the confusion matrix in Table 4.

Table 4. Confusion matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Evaluation metrics from confusion matrix are used to check the performance of model build. There are five metrics checked in this study, accuracy, precision, recall, specificity, and F1 score. Formula of each metric is as equations follow:

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

$$Accuracy = \frac{(TP + TN) \times 100\%}{TP + TN + FP + FN} \quad (21)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (22)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (23)$$

$$Specificty = \frac{TN}{TN + FP} \times 100\% \quad (24)$$

$$F1\ Score = \frac{2TP}{2TP + FP + FN} \quad (25)$$

III. RESULTS AND DISCUSSION

Now, the obtained data from preprocessing is ready to use for building Backpropagation model. In data preprocessing, number of observations is reduced to 769 data after eliminating the missing data. The distribution of dataset is shown in Figure 1.

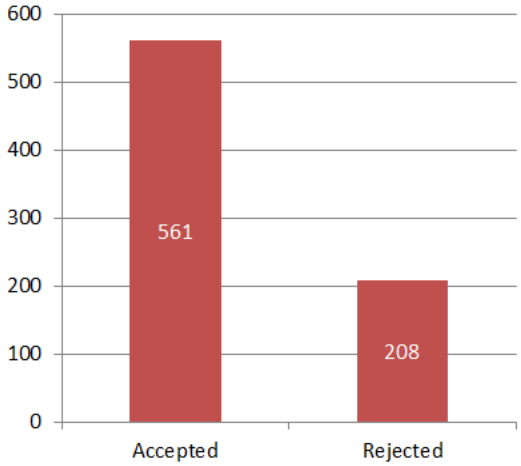

Figure 1. Distribution of loan status in the dataset

Furthermore, the result of feature selection has shown significant variables influenced prediction results. In Point Biserial Correlation Coefficient, Table 5 summarizes the result of estimates and p–value.

Every numerical variable has p–value bigger than 0.05 addressed with no correlation between these variables with the loan status.

Table 5. Point biserial correlation coefficient

	Estimates	*p*–value
Applicant Income	0.01267	0.7257
Co-applicant Income	0.04711	0.1919
Loan Amount	0.06509	0.0712

For Chi–square Test, the result is shown in Table 6. There are four variables that have p – value bigger than 0.05, namely Gender, Dependents, Education and Self Employed. These variables pointed to no correlation with loan status. Besides, there are also four variables have p–value less than 0.05, they are Married, Loan Term, Credit History, and Property Area. If p–value is less than 0.05, two variables have shown correlation.

Table 6. Chi-square test

	Estimates	*p*-value
Gender	1.6988	0.1924
Married	9.5649	0.001983
Dependents	2.6833	0.4431
Education	2.9874	0.08391
Self Employed	0.060117	0.8063
Loan Term	23.891	0.0132
Credit History	298.14	<2.2e-16
Property Area	12.883	0.001594

From feature selection conducted, for both Point Biserial correlation and Chi–square test, it shows that there are only four variables that significant towards the target variables. These four significant variables have p–value less than 0.05 and they are Married, Loan Term, Credit History, and Property Area. Two models of data are formed in this study, where one is four significant variables with loan status, and the second is every variable within dataset. Table 7 shows data models included variables.

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

Table 7. Data model for research

Data	Total Input	Variables
Model A	13	Gender, Married, Dependents, Education, Self Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Term, Credit History, Property Area: Rural, Property Area: Urban, Property Area: Semiurban.
Model B	4	Married, Loan Term, Credit History, Property Area

3.1 Building Backpropagation Model

In this section, building Backpropagation model is discussed. By applying the formula in (3) and the number of observations is 768, it is obtained that sample size is 284 data with 213 data for training and 71 data for testing. Important decision in building backpropagation model is to determine the number of units in hidden layer. Until today, there hasn't found formula for finding appropriate amount of hidden units theoretically. However, there are several rules of thumb to help researchers decide size of hidden unit, the rules are as follows [24]:

1. Hidden unit is 2/3 of size of input unit and output unit
2. Size of hidden unit is between size of input and output unit
3. Size of hidden unit is less than twice the size of the input layer

The number of hidden units for Backpropagation model by maximizing the rules of thumb for model A and model B are summarized in Table 8.

Table 8. Model design for backpropagation algorithm

Data	Model	Input Unit	Hidden Unit	Output Unit
Model A	A1	13	9	1
	A2		8	
	A3		7	
	A4		6	
Model B	B1	4	3	1
	B2		2	
	B3		1	

From Table 8, the appropriate number of hidden units for model A by applying first rule of thumb is 9 hidden units. For model B, by first rule of thumb, number of hidden units is 3. Additional model was created by reducing the number of hidden units. In addition, learning rate is also one of important hyperparameters in training model. Such a formula to obtain appropriate learning rate hasn't been found, therefore it's practical to trial and error. A traditional value to start is 0.1 and reduces the number by logarithmic scale [25]. After several trial errors, this research finally concludes the best learning rate that appropriate for model. For model A, with total of 4 models, this study uses a learning rate of 0.01 for training and testing. Summary of parameters of every model A is shown in Table 9.

Table 9. Parameter in model A

Model	Hidden Unit	Learning Rate	Activation Function
A1	9	0.01	Sigmoid
A2	8	0.01	Sigmoid
A3	7	0.01	Sigmoid
A4	6	0.01	Sigmoid

The learning rate for model B prediction is 0.1. The summary of model B parameter applied in this research is shown in Table 10.

Table 10. Parameter in model B

Model	Hidden Unit	Learning Rate	Activation Function
B1	3	0.1	Sigmoid
B2	2	0.1	Sigmoid
B3	1	0.1	Sigmoid

Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (147-158)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.27006

After generating model for prediction, then proceeds to test the model in classifying the loan status of applicants. One more step after testing is to check the performance of Backpropagation model in resolving loan status prediction by using evaluation matrix in confusion matrix.

3.2 Backpropagation Model Testing

In testing the Backpropagation model, the algorithm will produce output that is accurate with the actual output. Output of this research is number between 0 and 1 or a probability, that later will generate accepted or rejected. The probability less than 0.5 is determined to be rounded to 0 or interpreted as rejected loan. If probability is bigger than 0.5, the rounding value will be 1 and this output interpreted as an acceptable loan. The testing process uses 71 data with distribution shown in Figure 2.

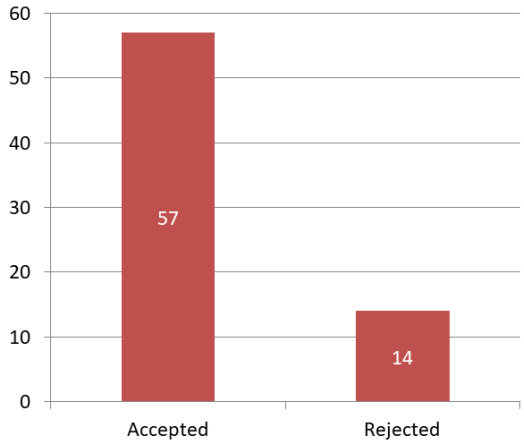

Figure 2. Distribution of loan status in testing data

This study is to predict output desired in binary classification, hence the performance assessment of each model is examined with a 2×2 confusion matrix. Evaluation metrics mentioned in previous section are accuracy, precision, recall, specificity, and F1 score. The proposed loan status classification is carried out with different data model and different architectures of Backpropagation. Performance of model A in classifying the loan status is summarized in Table 11.

Model A includes every variable from dataset, there are correlated variables and there are also variables that have no correlation with Loan Status. The lowest accuracy is model A4 with 80.28% and highest is model A3 with 94.37%. The sensitivity with highest result is from model A3 with 78.57%, then model A1 with 71.43%, and model A2 with 64.29%. As for the specificity, all models in model A obtained a very high number above 90%. Additionally, another performance metric is model A4 cannot be general optimal since unfortunate occurred in testing the model with A4.

Table 11. Summary of model A performance

Model	Architecture	Accuracy	Sensitivity	Specificity	Precision	F1 Score
A1	13 – 9 – 1	91.55%	71.43%	96.49%	83.33%	76.92%
A2	13 – 8 – 1	91.55%	64.29%	98.25%	90.00%	75.00%
A3	13 – 7 – 1	94.37%	78.57%	98.25%	91.67%	84.62%
A4	13 – 6 – 1	80.28%	0.00%	98.28%	0.00%	N/A

Table 12. Summary of model B performance

Model	Architecture	Accuracy	Sensitivity	Specificity	Precision	F1 Score
B1	4 – 3 – 1	94.37%	78.57%	98.25%	91.67%	84.62%
B2	4 – 2 – 1	94.37%	78.57%	98.25%	91.67%	84.62%
B3	4 – 1 – 1	94.37%	78.57%	98.25%	91.67%	84.62%

Performance model B is also summarized in Table 12 displayed above. All results of model B obtained 94.37% the accuracy and same values for other performance metrics. Model B used the 4 data variables that correlated with Loan Status. For model B, all Backpropagation models have sensitivity of 78.57%, specificity of 98.25%, precision of 91.67%, and F1 Score of 84.62%. Based on Table 11 and Table 12, the model with applied Backpropagation in predicting and classifying the loan status of applicants has good accuracy. The smallest accuracy was obtained from Model A4 with an architecture of 6 hidden nodes. Model A4 accurately classifies the loan status by 80.28%. The highest accuracy obtained is 94.37%.

From above Table 11 and Table 12, each model has obtained at least one backpropagation model with accuracy 94.37%. In addition, model B gives identical results in predicting loan status. All this is because the variables in model B consist of significant input variables towards loan status. For the case of the given dataset, based on the results of model A and model B, it can be said that the process of forming model B is more efficient because it requires a simple architecture and a smaller amount of data than in model A. Thus, it becomes important to analyze the correlation between predictor variables and target variables before the Backpropagation model is constructed.

For the limitation, this research was limited to not optimize the performance metrics of the model by making the variation of the parameter for example parameter of learning rate. In addition, the research was limited to the binary classification of loan applicant status. This matter is related to limited information could be obtained from dataset. The dataset used for this research has size less than one thousand and relatively small variable provided in the data.

IV. CONCLUSION

This study has applied the Backpropagation algorithm to predict the loan status with the used dataset from [15]. The good performance metrics of the model will help the financial institution to reduce credit risk.

The two simulation experiments, model A and model B, were presented. Model A is a research model in which all input variables exist in the dataset. Model A involves 13 independent inputs and consists of four model architectures distinguished by the number of hidden layers. Model B filters the predictor variables with those variables that show association with target variable being used for model B. As a result, model B involves only 4 independent inputs and has three model architectures.

The result of performance metrics of model A shows the best performance metrics are 94.37% accuracy, 78.57% sensitivity, 98.25% specificity, 91.67% precision, and 84.62% F1 score. The other simulation has the same with this result. For the given dataset, this means that analysis of correlation between predictor and target variable is important since can make the backpropagation model constructed efficiently.

By comparing these results with previous works, for example in [4] where author used Logistic Regression and Naïve Bayes classifier for loan prediction, the accuracy obtained 85.9% and 84.62% respectively, then it can be concluded that the backpropagation algorithm gives the better of performance than the both one.

Further study might use vary the parameter model for example learning rate to find the optimum of the performance metrics of the model. On the other hand, To find a more realistic result, it is important for further research by using historical dataset from financial institutions regarding credit applicants supposed to image a complex dataset that consists of more data variable and bigger size of data.

The result of this expected financial institutions to consider the use of Backpropagation algorithm in modeling predictive application before they proving credit to clients.

BIBLIOGRAPHY

[1] R. Indonesia, *Undang-Undang RI No. 10 Tahun 1998 tentang Perbankan*. 1998.

[2] A. Hornby and D. Lea, *Oxford Advanced Learner's Dictionary of Current English*, 10th editi. Oxford: Oxford University Press, 2020.

[3] S. Heffernan, *Modern Banking*, vol. 16, no. 3. 2016.

[4] Christabell, "Prediction of loan status using logistiscs regression model and naïve bayes classifier," President University, 2022.

[5] S. M. Fati, "Machine Learning-Based Prediction Model for Loan Status Approval," *J. Hunan Univ. Nat. Sci.*, vol. 48, no. 10, 2021, [Online]. Available: http://jonuns.com/index.php/journal/article/view/783.

[6] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 490–494, 2020, doi: 10.1109/ICESC48915.2020.9155614.

[7] C. Mason, J. Twomey, D. Wright, and L. Whitman, "Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression," *Res. High. Educ.*, vol. 59, no. 3, pp. 382–400, 2018, doi: 10.1007/s11162-017-9473-z.

[8] M. R. Romadhon and F. Kurniawan, "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," *3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021*, pp. 41–44, 2021, doi: 10.1109/EIConCIT50028.2021.9431845.

[9] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput. Sci.*, vol. 162, no. Itqm 2019, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.

[10] S. Hassanipour *et al.*, "Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis," *Injury*, vol. 50, no. 2, pp. 244–250, Feb. 2019, doi: 10.1016/j.injury.2019.01.007.

[11] A. Al Imran, M. N. Amin, and F. T. Johora, "Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning," in *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, Dec. 2018, pp. 1–6, doi: 10.1109/CIET.2018.8660844.

[12] L. Zajmi, F. Y. H. Ahmed, and A. A. Jaharadak, "Concepts, Methods, and Performances of Particle Swarm Optimization, Backpropagation, and Neural Networks," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, pp. 1–7, Sep. 2018, doi: 10.1155/2018/9547212.

[13] S. Setti and A. Wanto, "Analysis of Backpropagation Algorithm in Predicting the Most Number of Internet Users in the World," *J. Online Inform.*, vol. 3, no. 2, p. 110, 2019, doi: 10.15575/join.v3i2.205.

[14] K. K. Hiran, R. D. Doshi, and R. Jain, *Machine Learning Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications, 2021.

[15] A. K. Jana, "R-Machine-Learning," *GitHub, Inc.*, 2018. https://github.com/anup-jana/R-Machine-Learning/tree/master/R Scripts/Datasets.

[16] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. S.I.: Morgan Kaufmann, 2022.

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2019.

[18] T. U. Islam and M. Rizwan, "Comparison of correlation measures for nominal data," *Commun. Stat. - Simul. Comput.*, vol. 51, no. 3, pp. 698–714, Mar. 2022, doi: 10.1080/03610918.2020.1869984.

[19] T. Hailemeskel Abebe, "The Derivation and Choice of Appropriate Test Statistic (Z, t, F and Chi-Square Test) in Research Methodology," *Math. Lett.*, vol. 5, no. 3, pp. 33–40, 2019, doi: 10.11648/j.ml.20190503.11.

[20] C. Uakarn, K. Chaokromthong, and N. Sintao, "Sample Size Estimation using Yamane and Cochranand Krejcie and Morgan and Green Formulas and Cohen Statistical Power Analysis by G*Power and Comparisons," *APHEIT Int. J.*, vol. 10 No. 2, pp. 76–88, 2021, [Online].

Available: https://so04.tci-thaijo.org/index.php/ATI/article/view/254253/173847.

[21] A. P. Windarto *et al.*, *Jaringan Saraf Tiruan: Algoritma Prediksi dan Implementasi*. Yayasan Kita Menulis, 2020.

[22] J. Feng and S. Lu, "Performance Analysis of Various Activation Functions in Artificial Neural Networks," *J. Phys. Conf. Ser.*, vol. 1237, no. 2, 2019, doi: 10.1088/1742-6596/1237/2/022030.

[23] A. Jain, A. Fandago, and A. Kapoor, *TensorFlow Machine Learning Projects*. Packt Publishing, 2018.

[24] S. D. Desai, S. Giraddi, P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon, *Back-propagation neural network versus logistic regression in heart disease classification*, vol. 702. Springer Singapore, 2019.

[25] J. Brownlee, *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*. Machine Learning Mastery, 2018.